



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

UNIT – I - Fundamentals of Genomics and Proteomics– SBI1309

Organization of prokaryotic and eukaryotic genomes

Prokaryotic

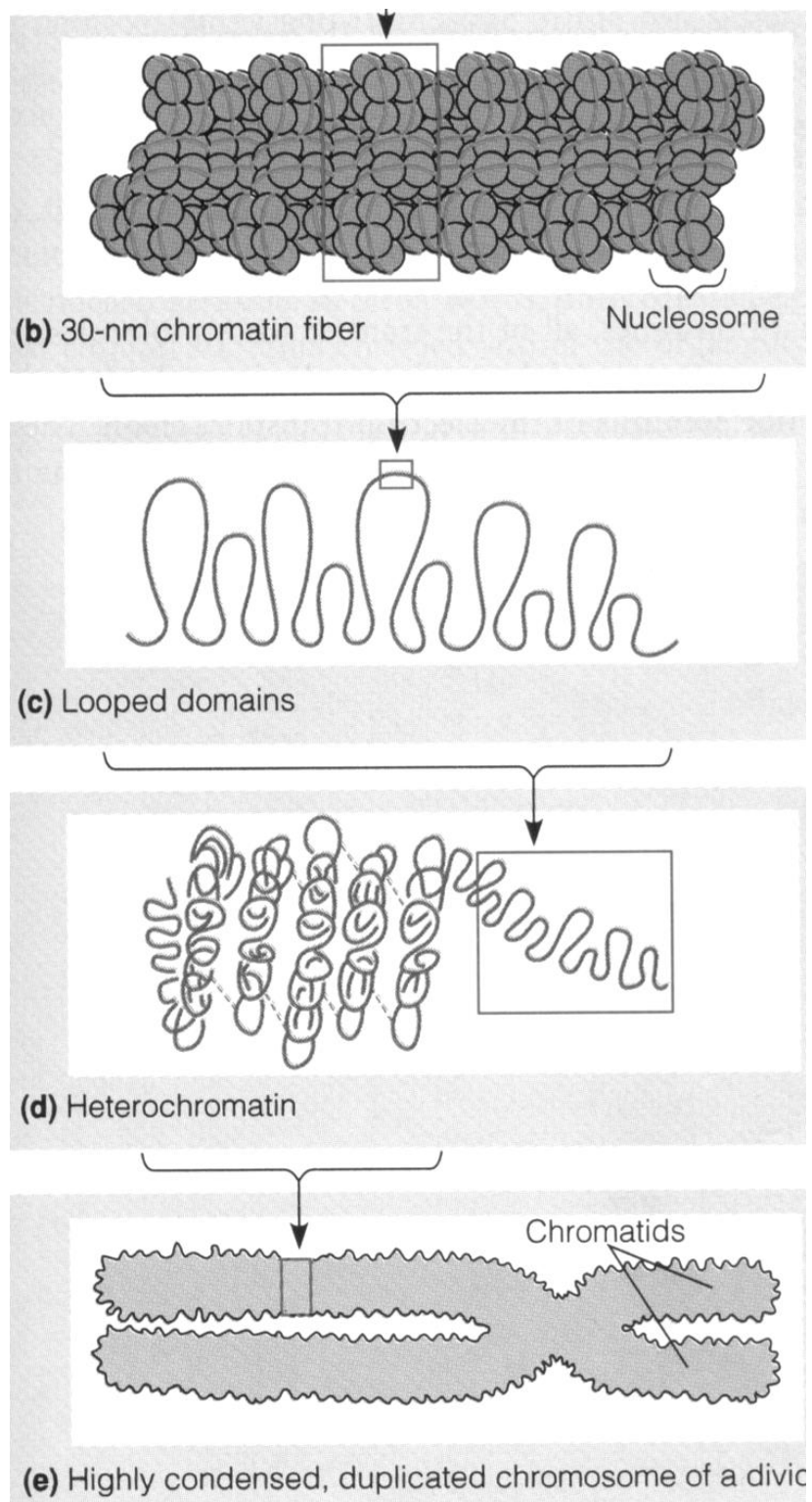
- Usually circular
- Smaller
- Found in the nucleoid region
- Less elaborately structured and folded

Eukaryotic

- Complexed with a large amount of protein to form chromatin
- Highly extended and tangled during interphase
- Found in the nucleus

The current model for progressive levels of DNA packing:

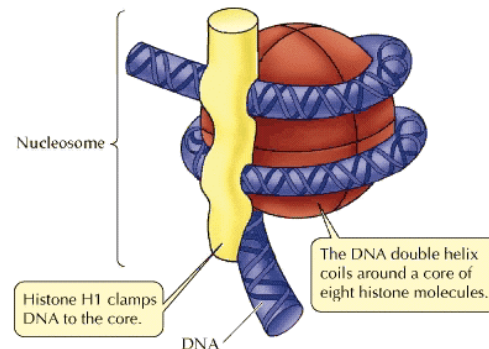
- Nucleosome → basic unit of DNA packing formed from DNA wound around a protein core that consists of 2 copies each of the 4 types of histone (H2A, H2B, H3, H4)]
- A 5th histone (H1) attaches near the bead when the chromatin undergoes the next level of packing
- 30 nm chromatin fiber → next level of packing; coil with 6 nucleosomes per turn
- the 30 nm chromatin forms looped domains, which are attached to a nonhistone protein scaffold (contains 20,000 – 100,000 base pairs)
- Looped domains attach to the inside of the nuclear envelope the 30 nm chromatin forms looped domains, which are attached to a nonhistone protein scaffold (contains 20,000 – 100,000 base pairs)



Histones influence folding in eukaryotic DNA.

- Histones → small proteins rich in basic amino acids that bind to DNA, forming chromatin

Contain a high proportion of positively charged amino acids which bind tightly to the negatively charged DNA



Heterochromatin

- Chromatin that remains highly condensed during interphase and is NOT actively transcribed

Euchromatin

- Chromatin that is less condensed during interphase and IS actively transcribed
- Becomes highly condensed during mitosis

Satellite DNA

→ highly repetitive DNA consisting of short unusual nucleotide sequences that are tandemly repeated 1000's of times

- It is found at the tips of chromosomes and the centromere

Its function is not known, perhaps it plays a structural role during chromosome replication and separation

Table 19.1 Types of Repetitive DNA

Tandemly Repetitive DNA (Satellite DNA)

Repeated units at a site are usually identical

Proportion of mammalian DNA:	10–15%
Length of each repeated unit:	1–10 base pairs
Total length of repetitive DNA per site, in base pairs:	
Regular satellite DNA	100,000–10 million
Minisatellite DNA	100–100,000
Microsatellite DNA	10–100

Interspersed Repetitive DNA

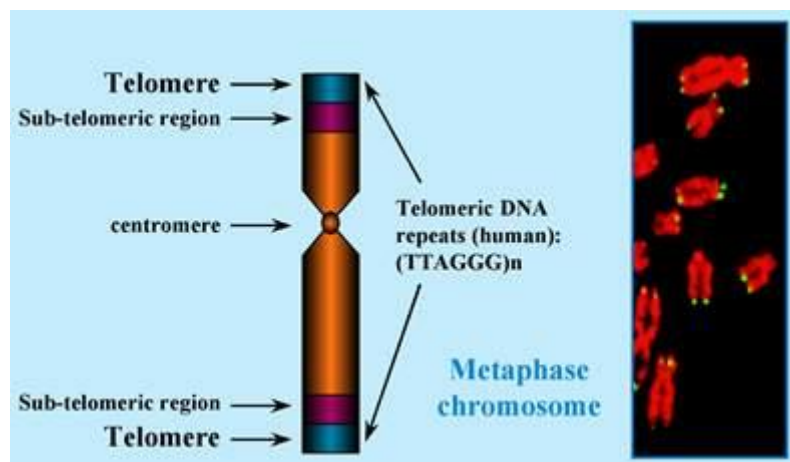
“Copies” are very similar but not identical

Proportion of mammalian DNA:	25–40%
Length of each repeated unit:	100–10,000 base pairs
Number of repetitions per genome:	10–1 million

Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Telomere→ series of short tandem repeats at the ends of eukaryotic chromosomes; prevents chromosomes from shortening with each replication cycle

Telomerase→ enzyme that periodically restores this repetitive sequence to the ends of DNA molecules



Genome Packaging in Prokaryotes: the Circular Chromosome of *E. coli*

***E. coli*: A Model Prokaryote**

Much of what is known about prokaryotic chromosome structure was derived from studies of *Escherichia coli*, a bacterium that lives in the human colon and is commonly used in laboratory cloning experiments. In the 1950s and 1960s, this bacterium became the model organism of choice for prokaryotic research when a group of scientists used phase-contrast microscopy and autoradiography to show that the essential genes of *E. coli* are encoded on a

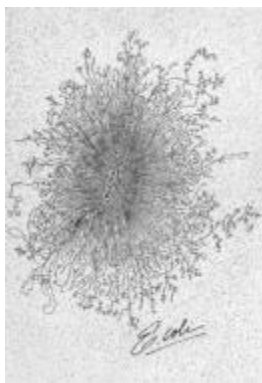
single circular chromosome packaged within the cell nucleoid (Mason & Powelson, 1956; Cairns, 1963).

Prokaryotic cells do not contain nuclei or other membrane-bound organelles. In fact, the word "prokaryote" literally means "before the nucleus." The nucleoid is simply the area of a prokaryotic cell in which the chromosomal DNA is located. This arrangement is not as simple as it sounds, however, especially considering that the *E. coli* chromosome is several orders of magnitude larger than the cell itself. So, if bacterial chromosomes are so huge, how can they fit comfortably inside a cell—much less in one small corner of the cell?

DNA Supercoiling

The answer to this question lies in DNA packaging. Whereas eukaryotes wrap their DNA around proteins called histones to help package the DNA into smaller spaces, most prokaryotes do not have histones (with the exception of those species in the domain Archaea). Thus, one way prokaryotes compress their DNA into smaller spaces is through supercoiling (Figure 1). Imagine twisting a rubber band so that it forms tiny coils. Now twist it even further, so that the original coils fold over one another and form a condensed ball. When this type of twisting happens to a bacterial genome, it is known as supercoiling. Genomes can be negatively supercoiled, meaning that the DNA is twisted in the opposite direction of the double helix, or positively supercoiled, meaning that the DNA is twisted in the same direction as the double helix. Most bacterial genomes are negatively supercoiled during normal growth.

Proteins Involved in Supercoiling



During the 1980s and 1990s, researchers discovered that multiple proteins act together to fold and condense prokaryotic DNA. In particular, one protein called HU, which is the most abundant protein in the nucleoid, works with an enzyme called topoisomerase I to bind DNA and introduce sharp bends in the chromosome, generating the tension necessary for negative supercoiling. Recent studies have also shown that other proteins, including integration host

factor (IHF), can bind to specific sequences within the genome and introduce additional bends (Rice *et al.*, 1996). The folded DNA is then organized into a variety of conformations (Sinden & Pettijohn, 1981) that are supercoiled and wound around tetramers of the HU protein, much like eukaryotic chromosomes are wrapped around histones (Murphy & Zimmerman, 1997).

Once the prokaryotic genome has been condensed, DNA topoisomerase I, DNA gyrase, and other proteins help maintain the supercoils. One of these maintenance proteins, H-NS, plays an active role in transcription by modulating the expression of the genes involved in the response to environmental stimuli. Another maintenance protein, factor for inversion stimulation (FIS), is abundant during exponential growth and regulates the expression of more than 231 genes, including DNA topoisomerase I (Bradley *et al.*, 2007).

Accessing Supercoiled Genes

Supercoiling explains how chromosomes fit into a small corner of the cell, but how do the proteins involved in replication and transcription access the thousands of genes in prokaryotic chromosomes when everything is packaged together so tightly? It has been determined that prokaryotic DNA replication occurs at a rate of 1,000 nucleotides per second, and prokaryotic transcription occurs at a rate of about 40 nucleotides per second (Lewin, 2007), so bacteria must have highly efficient methods of accessing their DNA strands. But how?

Researchers have noted that the nucleoid usually appears as an irregularly shaped mass within the prokaryotic cell, but it becomes spherical when the cell is treated with chemicals to inhibit transcription or translation. Moreover, during transcription, small regions of the chromosome can be seen to project from the nucleoid into the cytoplasm (i.e., the interior of the cell), where they unwind and associate with ribosomes, thus allowing easy access by various transcriptional proteins (Dürrenberger *et al.*, 1988). These projections are thought to explain the mysterious shape of nucleoids during active growth. When transcription is inhibited, however, the projections retreat into the nucleoid, forming the aforementioned spherical shape.

Because there is no nuclear membrane to separate prokaryotic DNA from the ribosomes within the cytoplasm, transcription and translation occur simultaneously in these organisms. This is strikingly different from eukaryotic chromosomes, which are confined to the membrane-bound nucleus during most of the cell cycle. In eukaryotes, transcription must be completed in the nucleus before the newly synthesized mRNA molecules can be transported to the cytoplasm to undergo translation into proteins.

Variations in Prokaryotic Genome Structure

Recently, it has become apparent that one size does not fit all when it comes to prokaryotic chromosome structure. While most prokaryotes, like *E. coli*, contain a single circular DNA molecule that makes up their entire genome, recent studies have indicated that some prokaryotes contain as many as four linear or circular chromosomes. For example, *Vibrio cholerae*, the bacteria that causes cholera, contains two circular chromosomes. One of these chromosomes contains the genes involved in metabolism and virulence, while the other contains the remaining essential genes (Trucksis *et al.*, 1998). An even more extreme example is provided by *Borrelia burgdorferi*, the bacterium that causes Lyme disease. This organism is transmitted through the bite of deer ticks (Figure 2), and it contains up to 11 copies of a single linear chromosome (Ferdows & Barbour, 1989). Unlike *E. coli*, *Borrelia* cannot supercoil its linear chromosomes into a tight ball within the nucleoid; rather, these strands are diffused throughout the cell.

Other organisms, such as *Bacillus subtilis*, form nucleoids that closely resemble those of *E. coli*, but they use different architectural proteins to do so. Furthermore, the DNA molecules of Archaea, a taxonomic domain composed of single-celled, nonbacterial prokaryotes that share many similarities with eukaryotes, can be negatively supercoiled, positively supercoiled, or not supercoiled at all. It is important to note that archaeans are the only group of prokaryotes that use eukaryote-like histones, rather than the architectural proteins described above, to condense their DNA molecules (Sandman *et al.*, 1990). The acquisition of histones by archaeans is thought to have paved the way for the evolution of larger and more complex eukaryotic cells.

Other DNA Differences Between Prokaryotes and Eukaryotes

Most prokaryotes reproduce asexually and are haploid, meaning that only a single copy of each gene is present. This makes it relatively easy to generate mutations in the lab and study the resulting phenotypes. By contrast, eukaryotes that reproduce sexually generally contain multiple chromosomes and are said to be diploid, because two copies of each gene exist—with one copy coming from each of an organism's parents.

Yet another difference between prokaryotes and eukaryotes is that prokaryotic cells often contain one or more plasmids (i.e., extrachromosomal DNA molecules that are either linear or circular). These pieces of DNA differ from chromosomes in that they are typically smaller and encode nonessential genes, such as those that aid growth in specific conditions or encode antibiotic resistance. *Borrelia*, for instance, contains more than 20 circular and linear

plasmids that encode genes responsible for infecting ticks and humans (Fraser *et al.*, 1997). Plasmids are often much smaller than chromosomes (i.e., less than 1,500 kilobases), and they replicate independently of the rest of the genome. However, some plasmids are capable of integrating into chromosomes or moving from cell to cell.

Perhaps due to the space constraints of packing so many essential genes onto a single chromosome, prokaryotes can be highly efficient in terms of genomic organization. Very little space is left between prokaryotic genes. As a result, noncoding sequences account for an average of 12% of the prokaryotic genome, as opposed to upwards of 98% of the genetic material in eukaryotes (Ahnert *et al.*, 2008). Furthermore, unlike eukaryotic chromosomes, most prokaryotic genomes are organized into polycistronic operons, or clusters of more than one coding region attached to a single promoter, separated by only a few base pairs. The proteins encoded by each operon often collaborate on a single task, such as the metabolism of a sugar into by-products that can be used for energy (Figure 3).

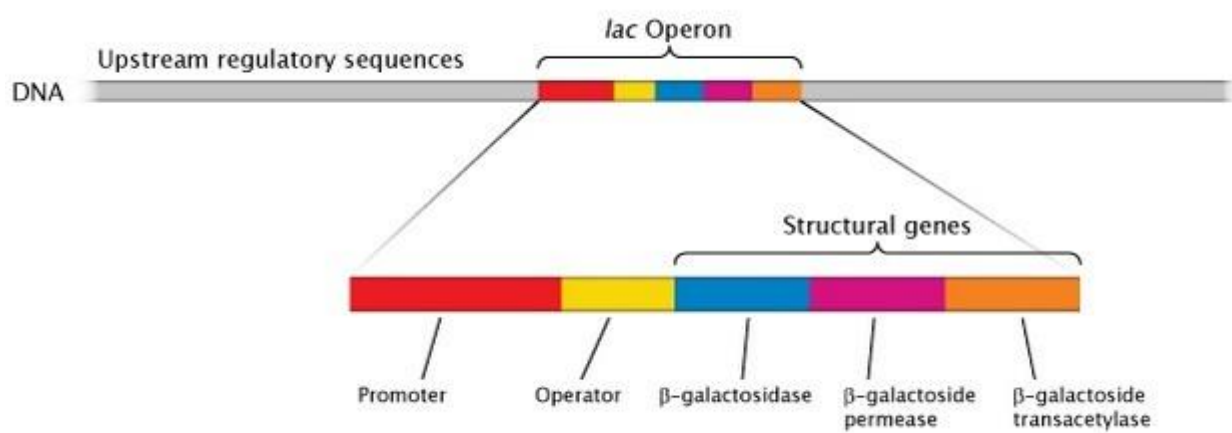


Figure 3: The prokaryotic *lac* operon.

Three structural genes code for proteins involved in lactose import and metabolism in bacteria. The genes are organized together in a cluster called the *lac* operon.

© 2013 Nature Education All rights reserved. 

Figure Detail

The organization of prokaryotic DNA therefore differs from that of eukaryotes in several important ways. The most notable difference is the condensation process that prokaryotic DNA molecules undergo in order to fit inside relatively small cells. Other differences, while not as dramatic, are summarized in Table 1.

Table 1: Prokaryotic versus Eukaryotic Chromosomes

Prokaryotic Chromosomes	Eukaryotic Chromosomes
<p>Many prokaryotes contain a single circular chromosome.</p> <p>Prokaryotic chromosomes are condensed in the nucleoid via DNA supercoiling and the binding of various architectural proteins.</p> <p>Because prokaryotic DNA can interact with the cytoplasm, transcription and translation occur simultaneously.</p> <p>Most prokaryotes contain only one copy of each gene (i.e., they are haploid).</p> <p>Nonessential prokaryotic genes are commonly encoded on extrachromosomal plasmids.</p> <p>Prokaryotic genomes are efficient and compact, containing little repetitive DNA.</p>	<p>Eukaryotes contain multiple linear chromosomes.</p> <p>Eukaryotic chromosomes are condensed in a membrane-bound nucleus via histones.</p> <p>In eukaryotes, transcription occurs in the nucleus, and translation occurs in the cytoplasm.</p> <p>Most eukaryotes contain two copies of each gene (i.e., they are diploid).</p> <p>Some eukaryotic genomes are organized into operons, but most are not.</p> <p>Extrachromosomal plasmids are not commonly present in eukaryotes.</p> <p>Eukaryotes contain large amounts of noncoding and repetitive DNA.</p>

Genome Projects

Genome projects are scientific endeavours that ultimately aim to determine the complete genome sequence of an organism (be it an animal, a plant, a fungus, a bacterium, an archaean, a protist or a virus) and to annotate protein-coding genes and other important genome-encoded features. The genome sequence of an organism includes the collective DNA sequences of each chromosome in the organism. For a bacterium containing a single chromosome, a genome project will aim to map the sequence of that chromosome. For the human species, whose genome includes 22 pairs of autosomes and 2 sex chromosomes, a complete genome sequence will involve 46 separate chromosome sequences.

What is genome science? (or genomics?) : It is the study of the structure, content, and evolution of genome.

Genome Projects

- Scientific endeavors that ultimately aim to determine the complete genome sequence of an organism.

- aim to increase understanding of genetics

- increase dimensionality of where and when

- Ultimate goal is establish an integrated web-based database and research interface

"Real" goals of genome projects

- To generate and order genomic and expressed gene sequences

- To identify and annotate genes of a genome

- To characterize DNA sequence diversity

Functional Genomics

- finding the meaning of the genome

- What are the biochemical, cellular, and/or physiological properties of each and every gene product of a genome?

- genome-scale gene expression profile

- genome-scale protein expression profile

- structural genomics, pharmacogenics

Comparative Genomics

Physical attributes of the DNA with locus information that may or may not include phenotypes.

Unit of genetic distance •centiMorgan

Physical Map

- An assembly of contiguous stretches of chromosomal DNA, aka "contigs".
(expressed in Kb)
- Work complementarily with genetic maps

Two general strategies to assemble contigs

- 1.Alignment of randomly isolated clones based on RFLP marker locations
- 2.Chromosomal walking

Cytological Maps

Cytological maps are the banding patterns observed through a microscope on stained chromosome spreads. Can be aligned with a physical map using insitu hybridization of cloned DNA fragments to the chromosomes a technique called FISH

Syntenly

A phenomenon in which local gene order along a chromosome tends to be conserved among different related species. Can be used to identify orthologs.

Chromosomal painting

A powerful method for determining syntenly

Orthologs

diverged by evolutionary split or speciation. We can use orthologs as anchoring landmarks for assembling a genome using syntenic segments from other species.

paralogs

diverged by gene duplication

Gene genealogy

An ultimate test for orthologous vs. paralogous relationships between homologs.

Assembling a genetic map

Smaller distance is much better than large distance because it can be prone to mutations/crossover.

Linkage group

Essentially a chromosome, which is ~100cM in animals; one crossover per chromosome per generation.

Genetic Map

The relative order of genetic markers in linkage groups in which the distance between markers is expressed as units of recombination.

Genetic markers

physical attributes of the DNA with locus information that may or may not include phenotypes. (With phenotypes: drosophila white mutation; without phenotypes: RFLP microsatellite markers)

Plant Genome projects

Since the publication in 2000 of the model *Arabidopsis thaliana* genome in the journal *Nature*, the number of genomes has steadily increased, peaking in 2012 with 13 publications (Fig. 1A). At this current trajectory there should be hundreds of plant genome publications over the next several years. Genome papers have been quite formulaic with a description of the assembly, gene numbers, repeats, WGDs, over and under-represented gene families, and finally, some aspect of novel biology, usually with a focus on transcription factors. Genomes have been published in 12 different journals with 38 of the 55 (69%) published genomes appearing in Nature journals (*Nature*, *Nature Genetics*, *Nature Biotech*, and *Nature Communications*); *Science* is second with six published genomes. As we see from the most recent publication of the *Capsella rubella* genome paper, the genome paper is shifting from a formulaic approach to a focus on how the genome elucidates novel biological aspects, such as the evolution of selfing to an outcrossing mating system (Slotte et al., 2013). The trend toward biology is quite positive and necessitated by demands for publication in high impact journals. However, the plant community is just at the beginning of exploring the diversity of plant genomes, and the rigor of the genome paper model with the associated in-depth exploration of genome features provides an essential foundation for the plant research community.

Table 1.

[View Full Table](#) | [Close Full View](#)

Published plant genomes.†

	Scientific name	Common name	Year	Type	Divisio
1	<i>Arabidopsis thaliana</i>	arabidopsis	2000	model	
2	<i>Oryza sativa</i>	rice	2002	crop	
3	<i>Oryza sativa</i>	rice	2002	crop	
4	<i>Oryza sativa</i>	rice	2005	crop	
5	<i>Populus trichocarpa</i>	black cottonwood	2006	crop	
6	<i>Vitis vinifera</i>	grape	2007	crop	
7	<i>Physcomitrella patens</i>	moss	2008	model	
8	<i>Vitis vinifera</i>	grape	2007	crop	
9	<i>Carica papaya</i>	papaya	2008	crop	
10	<i>Lotus japonicus</i>	lotus	2008	model	
11	<i>Sorghum bicolor</i>	sorghum	2009	crop	
12	<i>Cucumis sativus</i>	cucumber	2009	crop	
13	<i>Zea mays</i>	maize	2009	crop	
14	<i>Glycine max</i>	soybean	2010	crop	
15	<i>Brachypodium distachyon</i>	brachypodium	2010	model	
16	<i>Ricinus communis</i>	castor bean	2010	crop	
17	<i>Malus x domestica</i>	apple	2010	crop	
18	<i>Jatropha curcas</i>	jatropha	2010	crop	
19	<i>Theobroma cacao</i>	cocoa	2011	crop	
20	<i>Fragaria vesca</i>	strawberry	2011	crop	

One of the forces driving the rapid increase in fully sequenced plant genomes is the exponential decrease in cost and speed of genome sequencing fueled by high throughput DNA sequencing (Schatz et al., 2012). More than half of the published genomes have been sequenced entirely or partly using Sanger technology (Table 1), which provides long high quality ~1000 base pair (bp) reads. Sanger sequencing requires a cloning step and is time consuming with an expensive price tag, although the final result is usually high quality depending on the genome. When 454 came onto the scene in the early 2000s the cost of sequencing dropped an order of magnitude (US\$200K vs. US\$2 M) encouraging the emergence of consortia and funding for the sequencing of new genomes. Grape was the first genome published in 2007 using a combination of 454 and Sanger, and now there are at least 18 genomes that have used varying amounts of 454 sequence. Illumina and SOLiD sequencing changed the paradigm yet again providing very short reads (35–150 bp) at yet another order of magnitude lower cost than 454. Only two genome projects have used SOLiD for genome sequencing (strawberry and tomato); however, Illumina has played an exclusive

role in 12 genomes, and was used in combination with other technologies in another 17 genomes. Third generation sequencing technologies such as Pacific Bioscience (PacBio) promise long (>5 kb) single molecule reads that would greatly improve assembly of repeat rich plant genomes. PacBio long reads show great promise in resolving regions that the other sequencing technologies have problems with (skewed GC, homopolymers), but throughput and accuracy are two issues that still require attention. However, new sequencing technologies are only part of the future of plant genomes since tried and true methods, such as BACs (bacterial artificial chromosomes), are finding a place in hybrid sequencing approaches such as in the highly heterozygous pear genome (Wu et al., 2013).

Most of the plants chosen to be sequenced to date fit specific criteria such as size of research community, model organisms or economically important, small genome size, ploidy (diploid), availability of inbred lines (low heterozygosity), access to genetic and physical maps, expressed sequence tags (EST)/transcriptome and other genomic tools. Seventy-three percent (40) of the plant genome publications have been on crop species and some of these crop species double as model systems while several were sequenced purely for research such as *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Brachypodium distachyon*, *Physcomitrella patens* (moss), and *Selaginella moellendorffii* (spikemoss). Most (94%) genomes sequenced to date are Angiosperms, of which 36 are dicots and 16 are monocots, while only one gymnosperm (spruce), one bryophyte (moss), and one lycophyta (club-moss) have been sequenced (Table 1). Much of the early decisions about which genomes to sequence were driven by the Department of Energy Joint Genome Institute (JGI), which resulted in the publication and public availability (phytozome) of eleven of the highest quality plant genomes. The Beijing Genome Institute (BGI) has contributed consistently over the years starting with the rice genome, then ten additional genomes primarily based on Illumina technology, and now they have announced a large-scale plant genome sequencing project. However, a “1000 plant genome project” analogous to that in other communities has yet to emerge.

Plant Genomes both Large and Small

Plant genome sizes span several orders of magnitude from the carnivorous corkscrew plant (*Genlisea aurea*) at 63 megabases (Mb) to the rare Japanese *Paris japonica* at 148,000 Mb (Bennett and Leitch, 2011). The smallest published genome is the carnivorous bladderwort (*Utricularia gibba*) at 82 Mb, while the largest, the Norway Spruce (*Picea abies*), stands by itself at 19,600 Mb, compared to the second largest of maize at 2300 Mb and the overall median of 480 Mb (Table 1, Fig. 1B). Access to high quality reference genomes confirmed that long terminal repeats (LTRs) retrotransposons are a primarily driver of the dramatic size range in plants (El Baidouri and Panaud 2013). For the large barley genome (5100 Mb), where retrotransposons are abundant and more recently active, a powerful genomics resource was generated through an alternative “gene-ome” approach by anchoring a high quality genespace assembly on a deep physical map merged with high-density genetic maps (International Barley Genome Sequencing Consortium, 2012). In contrast, large gymnosperm genomes have highly diverged ancient repeats, which could make assembling these genomes tractable with current sequencing and assembly technologies (Kovach et al., 2010). The smallest reported conifer genome is the same size as maize and the median genome size is 9700 Mb, which is why a large push to sequence gymnosperms may have to wait for the next wave of sequencing technologies with increased read length and decreased price. As the community moves forward to choose the next round of genomes to sequence, the Kew Genome Size database will continue to provide a rich resource of non-model/non-crop species to investigate (Bennett and Leitch, 2011).

One measure of genome assembly quality is the contiguity or the length of contigs and scaffolds at which 50% of the assembly can be found; this is commonly referred to as N50. Sorghum, *Brachypodium distachyon*, soybean, and foxtail millet have the top four scaffold contiguities with 62.4, 59.3, 47.8, and 47.3 Mb respectively and all four were sequenced using Sanger as part of the JGI pipeline (Table 1). However, the genome with the ninth largest scaffold N50 is the tomato genome at 16 Mb, which was predominantly assembled using 454. Each scaffold is comprised of thousands of contigs and contig length generally drives the completeness and quality of the gene predictions. Not surprising, the 11 JGI assemblies based on Sanger have the top contig N50 ranging from 347 to 119 kilobases (kb), while the median contig N50 for all assemblies is 25.6 kb. Illumina based assemblies, primarily from BGI, have a similar median length (25.9 kb), which reflects their comprehensive strategy that makes use of different sized sequencing libraries. Another measure of a genome assembly is the amount of the genome captured in the assembly. Of the

published genomes, the median genome assembly captured 85% of the predicted genome size, which is usually estimated by flow cytometry or more recently by k-mer depth analysis. The remaining fraction of the genome not assembled generally represents the highly repetitive portion of the genome such as high copy number ribosomal repeats, centromeres, telomeres, and transposable elements. Therefore an average plant genome assembly captures 85% of the genome space in thousands of contigs with an N50 of 20 kb and tens of scaffolds with an N50 of 1 Mb.

Annotation of any genome, but particularly plant genomes, is difficult especially as the definition of what constitutes a gene continues to evolve. Many parts of the genome are ‘expressed’ in that RNAs are formed, but do not correspond to traditional genes in that they are not translated to a protein. However, most annotated plant genomes have between 20,000 and 94,000 genes with a median predicted gene count of 32,605 (Table 1, Fig. 1C). Differences between genomes most likely lies in the tools used for annotation and how relaxed the annotators were in calling genes as well as lineage-specific genes and gene family expansions. Genomes produced by next generation sequencing typically have smaller contig and scaffold sizes that complicate annotation as genes may not exist on single contigs but may be broken across contigs, thus inflating the number of annotated genes (e.g., pigeon pea, Varshney et al., 2012). Further complicating annotation is that there are many expressed non-coding RNAs that are functionally important (Eddy, 2001), but not considered genes in a traditional sense. Small RNA precursors are often not included in a genome annotation, but are important for plant development and silencing of TEs (Arikiti et al., 2013). Small RNAs and other non-coding RNAs are often annotated and curated separately from genome annotations in small, boutique databases. Long-term, however, one goal should be to combine these various sources of information into a single database/annotation making it easier for the biologist to pull together relevant information needed for forming hypotheses.

Plant genomes are packed, and often obese, with transposable elements (TEs) (Bennetzen 2000), which contain protein-coding sequences that are often annotated as genes. In rice, for instance, it was estimated that only 40,000 of the more than 55,000 annotated genes are really genes and that the other 10,000 to 15,000 are TEs—usually low copy TEs as high copy elements are relatively easy to find (Bennetzen et al., 2004). TEs include various families that move via copy-and-paste (class I) and cut-and-paste (class II) mechanisms. Copy-and-paste TEs can dramatically increase the size of a genome such as occurred in a relative of rice with a genome nearly two-fold larger than rice (Piegu et al., 2006). Transposon biology is an intriguing area of research and relies on relatively complete

genomes so that TEs are captured in sequence contigs and can be accurately annotated. Schemes for classification of TEs have been agreed on (Wicker et al., 2007), but annotation of non-LTR TEs is complicated by the lack of structural clues that allow routine ab initio prediction (El Baidouri and Panaud, 2013). Another complication is that in genomes produced by short read DNA sequencing technology, TEs are often missed in the assembly due to their repetitive nature. Genomes sequenced to date range from 3 to 85% repetitive sequence (Table 1; median 43%), with TEs, specifically cut-and-paste TEs (LTRs), comprising the majority of that sequence. Capturing and annotating these genomic components is important as it is becoming increasingly clear that TEs can be domesticated to function in gene regulation and as structural components of the genome.

Making Genomes “Functional”

One of the key take homes from the first 49 sequenced plant species is that we still have a lot to learn about the organization of genomes, function of genes, and how to characterize the non-coding space. Each new genome uncovers novel genes specific to a species, and a vast amount of non-coding space that requires methods for ab initio and functional annotation. One specific challenge is how we will leverage a growing number of high throughput technologies, otherwise referred to as “omics” approaches, to functionally annotate features of the plant genome. In this special issue of *The Plant Genome* we highlight several omics studies that have used high throughput approaches such as gen-omics (SNP detection), epigen-omics (methylation) metagen-omics (plant-fungal interactions), and ion-omics (element profiling) to refine our functional understanding of several key crop genomes (Eichten et al., 2013; Roorkiwal et al., 2013; Ruzicka et al., 2013; Ziegler et al., 2013). As we have seen through the model organism and human ENCODE projects, the layering of omics data exponentially increases the value of a reference genome (Celniker et al., 2009; ENCODE Project Consortium 2012).

While reference genomes provide a starting point, or platform for discovery in a specific species, it only captures a brief moment in the history of that species’ diversity and lacks the information content that would enable activities such as molecular breeding and phylogenetic analyses. Roorkiwal et al. (2013, this issue) describe the development of an Illumina BeadXpress SNP genotyping platform for two important crops in the developing world, pigeon pea and chickpea (Roorkiwal et al., 2013). Both pigeon pea and chickpea have lagged behind other crops in their genetic improvement due to a lack of genome and breeding resources that would enable such applications as marker assisted selection (MAS) and phylogenetic screens to identify genetic novelty in wild species. The development of an

Illumina BeadXpress SNP genotyping platform provides the opportunity to assess larger populations of plants with an adequate density of markers, which is ideal for breeding applications such as MAS and scans of diversity for disease and abiotic traits.

A prominent feature of plant genomes is their epigenetic landscape. The epigenome encompasses DNA methylation, histone modifications and other modifications not directly encoded in the genome. In general, DNA methylation is thought to mark permanent changes in the genome that must exist over the developmental lifetime of the plant, such as silencing transposable elements in embryonic tissue to protect the fidelity of the genome from transposition. Eichten et al. (2013, this issue) address the question of whether DNA methylation also specifies tissue types in maize. Using genome-wide array and sequencing technologies to assess DNA methylation and gene expression in two maize inbreds, B73 and Mo17, across four tissue types (leaf, immature tassel, embryo and endosperm), the authors find that there are more differentially methylated regions (DMRs) between maize inbreds than in the tissues they sampled (Eichten et al., 2013). The DMRs that were identified between tissue types did not correlate with subsequent expression changes suggesting the DMRs were not in fact functional in specifying tissue type. Despite other plants such as tomato that display tissue and developmentally regulated DMRs (Zhong et al., 2013), this may not be a general phenomenon in other species such as maize, which highlights the need to functionally define genomic elements in specific species.

Genetic screens are still the primary tool for functionally defining features of genomes. Mutant screens have been central in elucidating pathways, uncovering novel functionality of known genes, and allowing the discovery of novel non-coding features such as epigenetic regulation and small RNAs. Ziegler et al. (2013, this issue) describe a powerful high throughput mutant screen for elemental differences between field grown soy plants, which could be applied to any plant species with modestly sized seeds like soy (Ziegler et al., 2013). High throughput elemental profiling, or ionomics, is an emerging omics platform that provides a glimpse of a plant–soil environment and how that plant is accessing that environment. Ionomics screens have been powerful at detecting genetic factors controlling ion uptake but also have started to shed light on root architecture and morphology. Therefore, this high throughput screen, which is agnostic to plant species, has the potential to functionally characterize a plant organ, the root, which has traditionally been difficult to define genetically and molecularly in a field environment.

An almost uncharacterized area of plant biology is the complement of organisms that live mutually with plant communities, or the metagenome. In many plants, the acquisition of

inorganic minerals is facilitated by an active network of mycorrhizal associations between soil fungal species and plant roots. However, assessing how these fungal and plant species interact has been hampered by the fact that many fungal species cannot be cultured. The advent of high throughput sequencing has enabled an unprecedented opportunity to identify the genomic changes induced through these communal relationships. Ruzicka et al. (2013, this issue) use high throughput sequencing to characterize the transcriptomes of both the tomato genome and its arbuscular mycorrhizal fungal symbiont in the field (Ruzicka et al., 2013). Instead of culturing the symbiont, a metagenomic sequencing strategy was employed where RNA from a wild-type tomato plant and a mutant for reduced mycorrhizal colonization were sequenced and bioinformatically separated. This metagenomic analysis revealed a suite of genes for transport and cell wall remodeling required for the symbiotic relationship. Metagenomic sequencing will open up the opportunity to explore additional symbiotic relationships and further functionally characterize aspects of the genome that are not innate to the genome sequence.

Future Plant Genomes

The first ~50 plant genomes have provided a glimpse at the gene number, types and numbers of repeats, and how genomes grow and contract. However, we are just at the beginning of defining the functional aspects of plant genomes. To reach the goal of breeding better plants for future food, clothing, and energy, we will need to expand both the species sequenced, the number of species re-sequenced, and the type of omics data layered on genomes. Currently only one gymnosperm has been sequenced and no CAM (Crassulacean acid metabolism) photosynthetic plants have been sequenced. While we have come a long way over the past 13 yr since the publication of the *Arabidopsis* genome, we still have a long way to go before we will be able to engineer the plant of the future.

Insect Genome Projects

The importance of insects

Insects are the largest animal group in the world (75% of all species are insects) and are economically and ecologically extremely important, because most flowering plants depend on insects for their pollination. The honey bees alone, for example, pollinate 15-20 billion dollars worth of crop yearly in the United States.

But insects can also be severe agricultural pests, destroying 30% of our potential annual harvest, and can be vectors (intermediate pathogen carriers) for major diseases such as

malaria, sleeping sickness, Dengue fever, yellow fever, and elephantiasis (for a more complete list of insect-borne diseases see: <http://www.traveldocor.co.uk/insects.htm>). There are an estimated 300-500 million cases of malaria and up to 2.7 million deaths (mainly children) from malaria each year. But also the other diseases are equally serious. Elephantiasis, for example, disables 130 million people worldwide and 1.1 billion people, 17% of the world's population, are at risk of infection.

It will be clear, therefore, that insects can be both very beneficial and harmful (pests) and that a few insect species are hampering the welfare of many hundreds of millions of people especially in the developing countries of the third world. Thus, vast social benefits would be gained, if one selectively could reduce the populations of these pest insects.

Which insects have been sequenced so far?

Although much sequencing efforts have been focused on the twelve *Drosophila* species (1-3), other important insects have also been sequenced during the last eight years (Fig. 1). These insect species are: the malaria mosquito *Anopheles gambiae* (sequenced and published, ref. 4); the silkworm *Bombyx mori* (sequenced and published, refs. 5 and 6); the honey bee *Apis mellifera* (sequenced and published, ref. 7); the red flour beetle *Tribolium castaneum* (sequenced and published, ref. 8); the yellow fever mosquito *Aedes aegypti* (sequenced and published, ref. 9); the mosquito *Culex pipiens* (sequenced, but unpublished); three parasitic wasp species, belonging to the genus *Nasonia* (sequenced, but unpublished); the blood-sucking bug *Rhodnius prolixus* (in progress); the pea aphid *Acyrtosiphon pisum* (in progress); and the body louse *Pediculus humanus* (in progress).

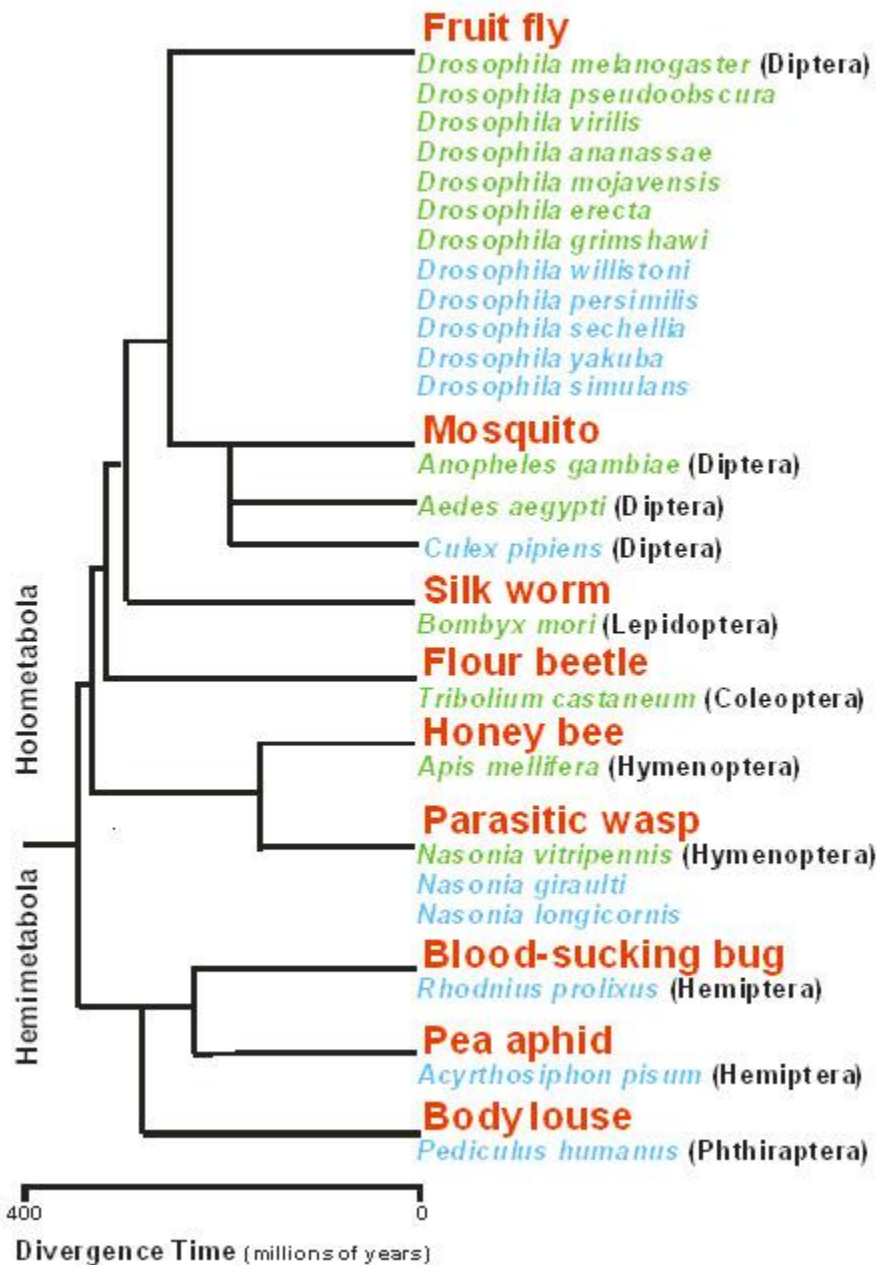


Fig. 1. Phylogenetic relationships of the insects, whose genomes have been sequenced. Green indicates genomes that have been fully sequenced (more than 8x coverage), blue indicates genomes, where the sequencing has not been completed (less than 8x coverage). The sequenced genomes cover about 350 million years of insect evolution.

These twenty-four species represent six different insect orders (Fig. 1) and also the two major evolutionary lineages of insects: Holometabola (insects with a complete metamorphosis) and Hemimetabola (insects having an incomplete metamorphosis). The sequenced genomes from these insects are goldmines, because they contain the information of all proteins and, thereby, of all biochemical and physiological processes that occur in an insect.

Why have these insects been selected for a genome project?

The twenty-four insect species have been selected for a genome project for different reasons. First, some insects are model organisms, such as the fruitfly *D. melanogaster*, which is a well-established model for geneticists and developmental and molecular biologists. The other *Drosophila* species have been sequenced to help with the interpretation of the *D. melanogaster* genome (2, 3).

A second class are the medically important insects that are vectors for serious diseases, such as malaria (the mosquito *A. gambiae*), yellow fever (the mosquito *A. aegypti*), elephantiasis (the mosquito *C. pipiens*), Chagas disease (the blood-sucking bug *R. prolixus*), and typhus (the body louse *P. humanus*).

A third class are the insects that are agriculturally important. To them belong agriculturally beneficial insects, such as the honey bee *A. mellifera*, which is a major pollinator of food plants and producer of honey, and the silkworm *B. mori*, which produces silk. In contrast to them, the red flour beetle *T. castaneum* (which destroys stored grain and many other dried and stored commodities for human consumption) and the pea aphid *A. pisum* (which causes severe damage to green food plants) are serious agricultural pests. The parasitic wasp *N. vitripennis* and the other two *Nasonia* species have been selected, because of their potentials for biological pest control (they lay eggs into a variety of agricultural pest insects).

Human Genome Project (HGP)

The **Human Genome Project (HGP)** was an international scientific research project with the goal of determining the sequence of chemical base pairs which make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and functional standpoint.^[1] It remains the world's largest collaborative biological project.^[2] The project was proposed and funded by the US government; planning started in 1984, got underway in 1990, and was declared complete in 2003. A parallel project was conducted outside of government by the Celera Corporation, or Celera Genomics, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in twenty universities and research centers in the United States, the United Kingdom, Japan, France, Germany, and China.^[3]

The Human Genome Project originally aimed to map the nucleotides contained in a human haploid reference genome (more than three billion). The "genome" of any given individual is unique; mapping "the human genome" involves sequencing multiple variations of each gene.^[4]

History

In May, 1985 Robert Sinsheimer organized a workshop to discuss sequencing the human genome,^[5] but for a number of reasons the NIH was uninterested in pursuing the proposal. The following March, the Santa Fe Workshop was organized by Charles DeLisi and David Smith of the Department of Energy's Office of Health and Environmental Research (OHER).^[6] At the same time Renato Dulbecco proposed whole genome sequencing in an essay in *Science*.^[7] James Watson followed two months later with a workshop held at the Cold Spring Harbor Laboratory.

The fact that the Santa Fe workshop was motivated and supported by a Federal Agency opened a path, albeit a difficult and tortuous one (Cook-Deegan),^[8] for converting the idea into public policy. In a memo to the Assistant Secretary for Energy Research (Alvin Trivelpiece), Charles DeLisi, who was then Director of OHER, outlined a broad plan for the project.^[9] This started a long and complex chain of events which led to approved reprogramming of funds that enabled OHER to launch the Project in 1986, and to recommend the first line item for the HGP, which was in President Regan's 1988 budget submission (Cook-Deegan),^[10] and ultimately approved by the Congress. Of particular importance in Congressional approval was the advocacy of Senator Peter Domenici, whom DeLisi had befriended.^[11] Domenici chaired the Senate Committee on Energy and Natural Resources, as well as the Budget Committee, both of which were key in the DOE budget process. Congress

added a comparable amount to the NIH budget, thereby beginning official funding by both agencies.

Dr. Alvin Trivelpiece sought and obtained the approval of DeLisi's proposal by Deputy Secretary William Flynn Martin. This chart^[12] was used in the spring of 1986 by Trivelpiece, then Director of the Office of Energy Research in the Department of Energy, to brief Martin and Under Secretary Joseph Salgado regarding his intention to reprogram \$4 million to initiate the project with the approval of Secretary Herrington. This reprogramming was followed by a line item budget of \$16 million in the Reagan Administration's 1987 budget submission to Congress.^[13] It subsequently passed both Houses. The Project was planned for 15 years.^[14]

Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.^[15]

In 1990, the two major funding agencies, DOE and NIH, developed a memorandum of understanding in order to coordinate plans and set the clock for the initiation of the Project to 1990.^[16] At that time, David Galas was Director of the renamed "Office of Biological and Environmental Research" in the U.S. Department of Energy's Office of Science and James Watson headed the NIH Genome Program. In 1993, Aristides Patrinos succeeded Galas and Francis Collins succeeded James Watson, assuming the role of overall Project Head as Director of the U.S. National Institutes of Health (NIH) National Center for Human Genome Research (which would later become the National Human Genome Research Institute). A working draft of the genome was announced in 2000 and the papers describing it were published in February 2001. A more complete draft was published in 2003, and genome "finishing" work continued for more than a decade.

The \$3-billion project was formally founded in 1990 by the US Department of Energy and the National Institutes of Health, and was expected to take 15 years.^[17] In addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Australia, China and myriad other spontaneous relationships.^[18]

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by U.S. President Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000).^[19] This first available rough draft assembly of the genome was completed by the Genome Bioinformatics Group at the University of California, Santa Cruz, primarily led by then graduate student Jim Kent. Ongoing sequencing led to the announcement of the essentially

complete genome on April 14, 2003, two years earlier than planned.^{[20][21]} In May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in *Nature*.^[22]

State of completion

The project did not aim to sequence all the DNA found in human cells. It sequenced only "euchromatic" regions of the genome, which make up about 90% of the genome. The other regions, called "heterochromatic" are found in centromeres and telomeres, and were not sequenced under the project.^[23]

The Human Genome Project was declared complete in April 2003. An initial rough draft of the human genome was available in June 2000 and by February 2001 a working draft had been completed and published followed by the final sequencing mapping of the human genome on April 14, 2003. Although this was reported to be 99% of the euchromatic human genome with 99.99% accuracy a major quality assessment of the human genome sequence was published on May 27, 2004 indicating over 92% of sampling exceeded 99.99% accuracy which was within the intended goal.^[24] Further analyses and papers on the HGP continue to occur.^[25]

What are the overall goals of the HGP?

The Human Genome Project has several goals, which include *mapping*, *sequencing*, and *identifying* genes, *storing* and *analyzing* data, and *addressing* the ethical, legal, and social issues (ELSI) that may arise from availability of personal genetic information. *Mapping* is the construction of a series of chromosome descriptions that depict the position and spacing of genes, which are on the DNA of chromosomes. ***The ultimate goal of the Human Genome Project is to decode, letter by letter, the exact sequence of all 3.2 billion nucleotide bases that make up the human genome.*** This means constructing *detailed genetic and physical maps of the human genome*. Besides determining the complete nucleotide sequence of human DNA, this includes locating the genes within the human genome. The HGP agenda also includes analyzing the genomes of several other organisms (including *E. coli*, the fruit fly, and the laboratory mouse) that are used extensively in research laboratories as model systems. Studying the genetic makeup of non-human organisms will help in understanding and deciphering the human genome. Although in recent months the leaders of the HGP announced that a “working draft” of the human Genome has been completed, the hope is to have a complete, error-free, final draft by 2003—coincidentally, the 50th anniversary of the discovery of DNA's molecular structure.

Summary of basic HGP goals:

- *Identify* all estimated 50,000-100,000 genes in human DNA
- *Determine sequence* of 3 billion chemical bases that make up human DNA
 - ***Human DNA sequence goals:***
 - Achieve *coverage* of at least 90% of Genome in *working draft* by the end of 2001—(moved up to spring 2000) - *Goal Reached* -
 - *Finish one-third* of the human Genome sequence by end of 2001
 - *Finish complete* human Genome sequence by end of 2003
 - Make sequence totally and freely accessible
- Create bioinformatics tools – Develop databases and analysis algorithms
- Store information in databases
- Develop faster, more efficient *sequencing technologies*
- Identify genes and coding regions – Develop efficient in-vitro or in-silico methods
- Develop tools for *data analysis*
- Map genomes of select *non-human* organisms
- Sequence other model organisms – Bacteria, yeast, fruit fly, worm, mouse
- Address *ethical, legal, and social issues* (ELSI) that may arise from project

Goals and Completion

Area	Goal	Achieved	Date Achieved
Genetic Map	2- to 5-cM resolution map (600 – 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < \$0.25 per finished base	Sequence >1,400 Mb/year at <\$0.09 per finished base	November 2002
Human Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene Identification	Full-length human cDNAs	15,000 full-length human cDNAs	March 2003
Model Organisms	Complete genome sequences of <i>E. coli</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i>	Finished genome sequences of <i>E. coli</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , plus whole-genome drafts of several others, including <i>C. briggsae</i> , <i>D. pseudoobscura</i> , mouse and rat	April 2003

Applications of Human Genome Project:

1. Better understanding of Polygenic disorders: The single gene disorders such as Cystic fibrosis, Sickle cell anemia are known. But many of the diseases such as Cancer, Hypertension are polygenic in nature. Sequencing of such genes helps us to better evaluate the disease giving more patient specific and friendly treatment.
2. Improvement in Gene Therapy: Genome sequencing helps in better provision of Gene therapy which is in its preliminary stage. This helps in effective treatment of genetic diseases.
3. Well elucidated Human genome sequence helps in improved diagnosis of many genetic disorders.
4. Development of Pharmacogenomics- Specialization in this field helps to know the individual genetic makeup thereby providing more personalized treatment.
5. Better cure of psychiatric disorders: Genes responsible for behavioral and psychiatric diseases can be better understood and treated.
6. An important application of HGP is better understanding of Mutations concept.

7. Better understanding of Developmental biology - Evolution from eggs to adults.
8. Human genome data also helps in development of Biotechnology in various spheres.

Findings

Large variation in GC content – Correlated with repeat content and gene density

- CpG dinucleotides are surprisingly rare – But CpG islands correlated with gene density
- Recombination rates are uneven – More recombination further from centromeres
- About 50% of genome is repeats – SINEs, LINEs, LTR retrotransposons, transposons

Mutation rates are uneven – Genome has more GC than equilibrium

- Differences between the sexes – Males mutate more but recombine less
- Many segmental duplications – 1–200 kb copied within or across chromosomes
- Estimated around 30,000 human genes – Unevenly distributed across chromosomes

MAJOR HIGHLIGHTS OF HGP :

1. Approximately 90pc of Human Genome was sequenced and the cause for underlying genetic disorders have been depicted.
2. The remaining 10pc is located at the end of chromosomes or at telomeres.
3. The human genome consisted of 3200 billion base pairs of which Gene and Gene Related sequences hold 1200 base pairs while Intergenic DNA contributed 2000 base pairs.
4. Proteins contribute 1.1-1.4 pc
5. Approximately 25% of the genome is composed of introns which appear as repeating units with no known functions.
6. Protein Coding Genes- 30000-40000
7. An average gene consists of 3000 bases. Dystrophin is the largest human gene with 2.4 million bases.
8. Chromosome 1 is the largest and contains 2968 genes while Chromosome Y is the smallest.
9. Genetic sequences that are associated with diseases like breast cancer, deafness, muscle diseases, blindness were sequenced and reported.
10. Repeated sequences constituted 50% of the genome.
11. 97% of the human genome has unknown functions
12. More than 3 million Single Nucleotide Polymorphisms have been identified.
13. Human DNA is 98% identical to Chimpanzees.
14. About 200 genes of human genome are found in bacteria too.

What is Next?

- Find all human genes – Only ~15,000 have yet been confirmed
- Identify effects of genetic variation – Understand diseases, healthy differences
- Understand non-coding regions – Chromosome structure, control mechanisms?
- Model human being as system – Within cells and as a whole organism

What were some of the ethical, legal, and social implications addressed by the Human Genome Project?

The Ethical, Legal, and Social Implications (ELSI) program was founded in 1990 as an integral part of the Human Genome Project. The mission of the ELSI program was to identify and address issues raised by genomic research that would affect individuals, families, and society. A percentage of the Human Genome Project budget at the National Institutes of Health and the U.S. Department of Energy was devoted to ELSI research.

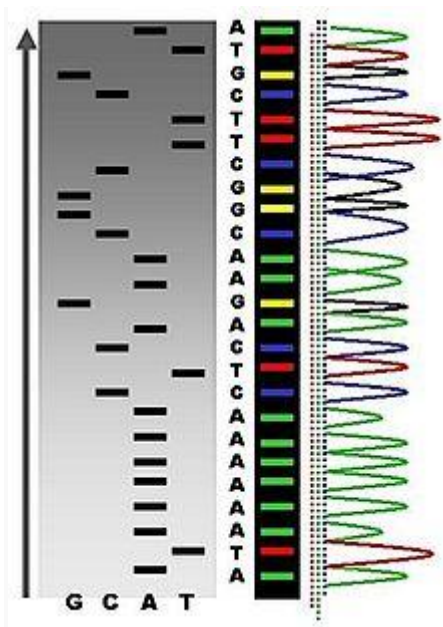
The ELSI program focused on the possible consequences of genomic research in four main areas:

- Privacy and fairness in the use of genetic information, including the potential for genetic discrimination in employment and insurance.
- The integration of new genetic technologies, such as genetic testing, into the practice of clinical medicine.
- Ethical issues surrounding the design and conduct of genetic research with people, including the process of informed consent.
- The education of healthcare professionals, policy makers, students, and the public about genetics and the complex issues that result from genomic research.

DNA sequencing

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as medical diagnosis, biotechnology, forensic biology, virology and biological systematics. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species.



An example of the results of automated chain-termination DNA sequencing.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of fluorescence-based sequencing methods with a DNA sequencer,^[1] DNA sequencing has become easier and orders of magnitude faster.

Uses of sequencing

DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e. clusters of genes or operons), full chromosomes or entire genomes. Sequencing provides the order of individual nucleotides present in molecules of DNA or RNA isolated from animals, plants, bacteria, archaea, or virtually any other source of genetic information.

This information is useful to various fields of biology and other sciences, medicine, forensics, and other areas of study.

Molecular biology

Sequencing is used in molecular biology to study genomes and the proteins they encode. Information obtained using sequencing allows researchers to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.

Evolutionary biology

Since DNA is an informative macromolecule in terms of transmission from one generation to another, DNA sequencing is used in evolutionary biology to study how different organisms are related and how they evolved.

Metagenomics

The field of metagenomics involves identification of organisms present in a body of water, sewage, dirt, debris filtered from the air, or swab samples from organisms. Knowing which organisms are present in a particular environment is critical to research in ecology, epidemiology, microbiology, and other fields. Sequencing enables researchers to determine which types of microbes may be present in a microbiome, for example.

Medicine

Medical technicians may sequence genes (or, theoretically, full genomes) from patients to determine if there is risk of genetic diseases. This is a form of genetic testing, though some genetic tests may not involve DNA sequencing.

Forensics

DNA sequencing may be used along with DNA profiling methods for forensic identification and paternity testing.

History

Deoxyribonucleic acid (DNA) was first discovered and isolated by Friedrich Miescher in 1869, but it remained understudied for many decades because proteins, rather than DNA, were thought to hold the genetic blueprint to life. This situation changed after 1944 as a result of some experiments by Oswald Avery, Colin MacLeod, and Maclyn McCarty demonstrated that purified DNA could change one strain of bacteria into another type. This was the first time that DNA was shown capable of transforming the properties of cells.

In 1953 James Watson and Francis Crick put forward their double-helix model of DNA which depicted DNA being made up of two strands of nucleotides coiled around each other, linked together by hydrogen bonds, in a spiral configuration. Each strand they argued was composed of four complementary nucleotides: adenine (A), cytosine (C), guanine (G) and

thymine (T) and was oriented in opposite directions. Such a structure they proposed allowed each strand to reconstruct the other and was central to the passing on of hereditary information between generations.^[7]

The foundation for sequencing DNA was first laid by the work of Fred Sanger who by 1955 had completed the sequence of all the amino acids in insulin, a small protein secreted by the pancreas. This provided the first conclusive evidence that proteins were chemical entities with a specific molecular pattern rather than a random mixture of material suspended in fluid. Sanger's success in sequencing insulin greatly electrified x-ray crystallographers, including Watson and Crick who by now were trying to understand how DNA directed the formation of proteins within a cell. Soon after attending a series of lectures given by Fred Sanger in October 1954, Crick began to develop a theory which argued that the arrangement of nucleotides in DNA determined the sequence of amino acids in proteins which in turn helped determine the function of a protein. He published this theory in 1958

RNA sequencing

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), in 1972^[9] and 1976.^[10]

Early DNA sequencing methods

The first method for determining DNA sequences involved a location-specific primer extension strategy established by Ray Wu at Cornell University in 1970.^[11] DNA polymerase catalysis and specific nucleotide labeling, both of which figure prominently in current sequencing schemes, were used to sequence the cohesive ends of lambda phage DNA.^{[12][13][14]} Between 1970 and 1973, Wu, R Padmanabhan and colleagues demonstrated that this method can be employed to determine any DNA sequence using synthetic location-specific primers.^{[15][16][17]} Frederick Sanger then adopted this primer-extension strategy to develop more rapid DNA sequencing methods at the MRC Centre, Cambridge, UK and published a method for "DNA sequencing with chain-terminating inhibitors" in 1977.^[18] Walter Gilbert and Allan Maxam at Harvard also developed sequencing methods, including one for "DNA sequencing by chemical degradation".^{[19][20]} In 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.^[21] Advancements in sequencing were aided by the concurrent development of recombinant DNA technology, allowing DNA samples to be isolated from sources other than viruses.

Sequencing of full genomes

The first full DNA genome to be sequenced was that of bacteriophage ϕ X174 in 1977.^[22] Medical Research Council scientists deciphered the complete DNA sequence of the Epstein-Barr virus in 1984, finding it contained 172,282 nucleotides. Completion of the sequence marked a significant turning point in DNA sequencing because it was achieved with no prior genetic profile knowledge of the virus.^[23]

A non-radioactive method for transferring the DNA molecules of sequencing reaction mixtures onto an immobilizing matrix during electrophoresis was developed by Pohl and co-workers in the early 1980s.^{[24][25]} Followed by the commercialization of the DNA sequencer "Direct-Blotting-Electrophoresis-System GATC 1500" by GATC Biotech, which was intensively used in the framework of the EU genome-sequencing programme, the complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome II.^[26] Leroy E. Hood's laboratory at the California Institute of Technology announced the first semi-automated DNA sequencing machine in 1986.^[27] This was followed by Applied Biosystems' marketing of the first fully automated sequencing machine, the ABI 370, in 1987 and by Dupont's Genesis 2000^[28] which used a novel fluorescent labeling technique enabling all four dideoxynucleotides to be identified in a single lane. By 1990, the U.S. National Institutes of Health (NIH) had begun large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* at a cost of US\$0.75 per base. Meanwhile, sequencing of human cDNA sequences called expressed sequence tags began in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.^[29] In 1995, Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) published the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*. The circular chromosome contains 1,830,137 bases and its publication in the journal Science^[30] marked the first published use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

By 2001, shotgun sequencing methods had been used to produce a draft sequence of the human genome

Next-generation sequencing methods

Several new methods for DNA sequencing were developed in the mid to late 1990s and were implemented in commercial DNA sequencers by the year 2000.

On October 26, 1990, Roger Tsien, Pepi Ross, Margaret Fahnestock and Allan J Johnston filed a patent describing stepwise ("base-by-base") sequencing with removable 3' blockers on DNA arrays (blots and single DNA molecules).^[33] In 1996, Pål Nyrén and his

student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm published their method of pyrosequencing.^[34]

On April 1, 1997, Pascal Mayer and Laurent Farinelli submitted patents to the World Intellectual Property Organization describing DNA colony sequencing.^[35] The DNA sample preparation and random surface-PCR arraying methods described in this patent, coupled to Roger Tsien et al.'s "base-by-base" sequencing method, is now implemented in Illumina's Hi-Seq genome sequencers.

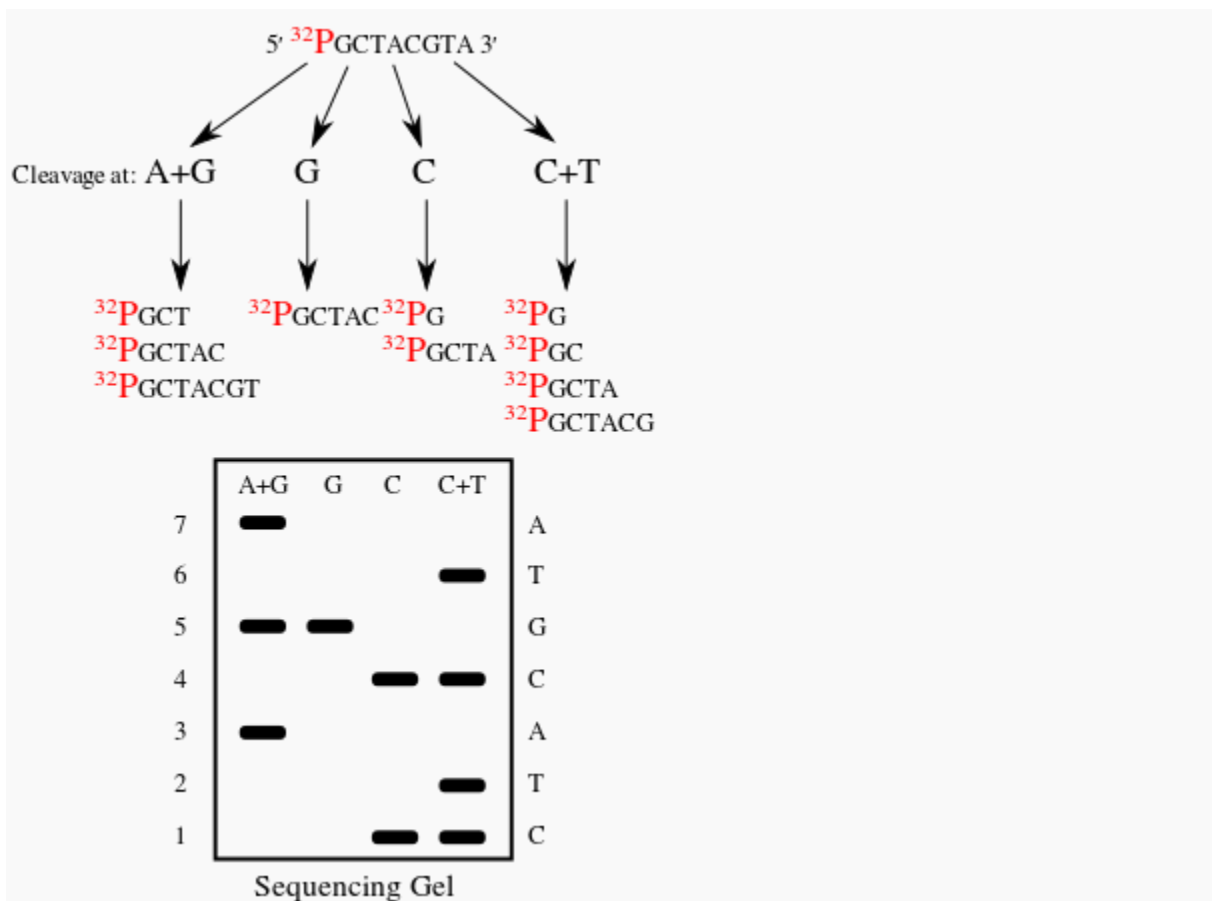
Lynx Therapeutics published and marketed "Massively parallel signature sequencing", or MPSS, in 2000. This method incorporated a parallelized, adapter/ligation-mediated, bead-based sequencing technology and served as the first commercially available "next-generation" sequencing method, though no DNA sequencers were sold to independent laboratories.^[36]

In 2004, 454 Life Sciences marketed a parallelized version of pyrosequencing.^[37] The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of the new generation of sequencing technologies, after MPSS.^[38]

The large quantities of data produced by DNA sequencing have also required development of new methods and programs for sequence analysis. Phil Green and Brent Ewing of the University of Washington described their phred quality score for sequencer data analysis in 1998

Maxam–Gilbert sequencing

Maxam–Gilbert sequencing is a method of DNA sequencing developed by Allan Maxam and Walter Gilbert in 1976–1977. This method is based on nucleobase-specific partial chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides.^[1]



An example Maxam–Gilbert sequencing reaction. Cleaving the same tagged segment of DNA at different points yields tagged fragments of different sizes. The fragments may then be separated by gel electrophoresis.

Maxam–Gilbert sequencing was the first widely adopted method for DNA sequencing, and, along with the Sanger dideoxy method, represents the first generation of DNA sequencing methods. Maxam–Gilbert sequencing is no longer in widespread use, having been supplanted by next-generation sequencing.

Procedure

Maxam–Gilbert sequencing requires radioactive labeling at one 5' end of the DNA fragment to be sequenced (typically by a kinase reaction using gamma-³²P ATP) and purification of the DNA. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are hydrolysed using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the reaction of thymine for the C-only reaction. The modified DNAs may then be cleaved by hot piperidine; (CH₂)₅NH at the position of the modified base. The concentration of the

modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule.

The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each showing the location of identical radiolabeled DNA molecules. From presence and absence of certain fragments the sequence may be inferred

Sanger sequencing

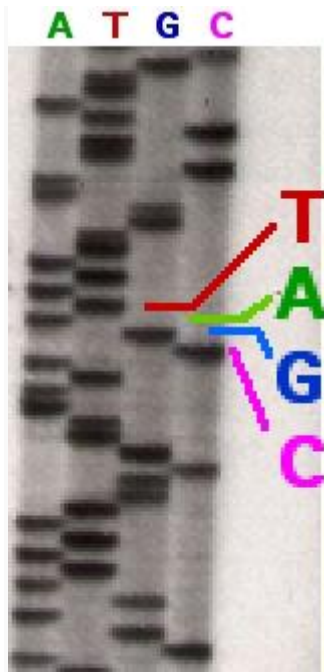
Sanger sequencing is a method of **DNA sequencing** based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.^{[1][2]} Developed by Frederick Sanger and colleagues in 1977, it was the most widely used sequencing method for approximately 25 years. More recently, Sanger sequencing has been supplanted by "Next-Gen" sequencing methods, especially for large-scale, automated genome analyses. However, the Sanger method remains in wide use, for smaller-scale projects, validation of Next-Gen results and for obtaining especially long contiguous DNA sequence reads (>500 nucleotides).

Method

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleosidetriphosphates (dNTPs), and modified di-deoxynucleosidetriphosphates (ddNTPs), the latter of which terminate DNA strand elongation. These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. The ddNTPs may be radioactively or fluorescently labeled for detection in automated sequencing machines.

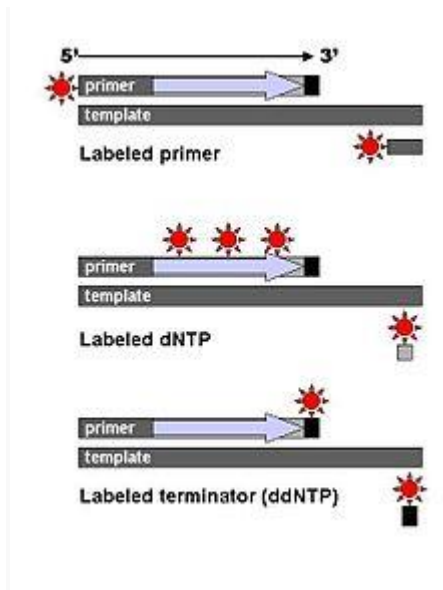
The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP), while the other added nucleotides are ordinary ones. The dideoxynucleotide is added in approximately 100-fold excess of the corresponding deoxynucleotide (e.g. 0.005mM dATP : 0.5mM ddATP) allowing for enough fragments to be produced while still transcribing the complete sequence.^[2] Putting it in a more sensible order, four separate reactions are needed in this process to test all four ddNTPs. Following rounds of template DNA extension

from the bound primer, the resulting DNA fragments are heat denatured and separated by size using gel electrophoresis. In the original publication of 1977,^[2] the formation of base-paired loops of ssDNA was a cause of serious difficulty in resolving bands at some locations. This is frequently performed using a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C). The DNA bands may then be visualized by autoradiography or UV light and the DNA sequence can be directly read off the X-ray film or gel image.



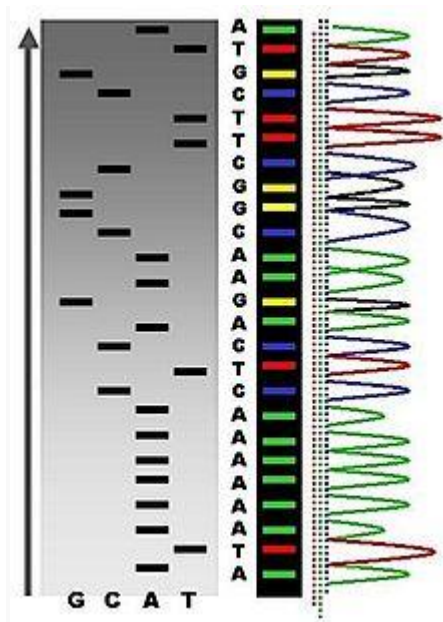
Part of a radioactively labelled sequencing gel

In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes, from bottom to top, are then used to read the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers^{[3][4]} of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.

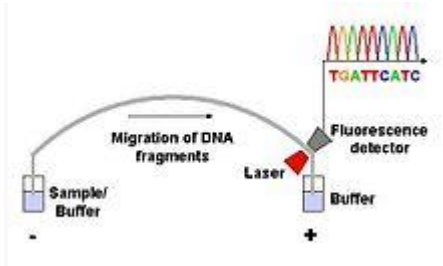


Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for

sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



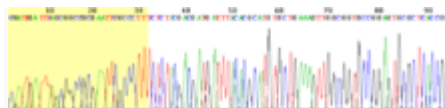
Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis (see figure to the left).

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch. Batch runs may occur up to 24 times a day. DNA sequencers separate strands by size (or length) using capillary electrophoresis, they detect and record dye fluorescence, and output data as fluorescent peak trace chromatograms. Sequencing reactions (thermocycling and labelling), cleanup and re-suspension of samples in a buffer solution are performed separately, before loading samples onto the sequencer. A number of commercial

and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (which are generally located at the ends of the sequence). The accuracy of such algorithms is inferior to visual examination by a human operator, but is adequate for automated processing of large sequence data sets.

Challenges

Common challenges of DNA sequencing with the Sanger method include poor quality in the first 15-40 bases of the sequence due to primer binding and deteriorating quality of sequencing traces after 700-900 bases. Base calling software such as Phred typically provides an estimate of quality to aid in trimming of low-quality regions of sequences.^{[5][6]}

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and next-generation sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.^{[7][8]}

Current methods can directly sequence only relatively short (300-1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide.

Microfluidic Sanger sequencing

Microfluidic Sanger sequencing is a lab-on-a-chip application for DNA sequencing, in which the Sanger sequencing steps (thermal cycling, sample purification, and capillary electrophoresis) are integrated on a wafer-scale chip using nanoliter-scale sample volumes. This technology generates long and accurate sequence reads, while obviating many of the significant shortcomings of the conventional Sanger method (e.g. high consumption of expensive reagents, reliance on expensive equipment, personnel-intensive manipulations, etc.) by integrating and automating the Sanger sequencing steps.

In its modern inception, high-throughput genome sequencing involves fragmenting the genome into small single-stranded pieces, followed by amplification of the fragments by Polymerase Chain Reaction (PCR). Adopting the Sanger method, each DNA fragment is irreversibly terminated with the incorporation of a fluorescently labeled dideoxy chain-terminating nucleotide, thereby producing a DNA “ladder” of fragments that each differ in length by one base and bear a base-specific fluorescent label at the terminal base. Amplified

base ladders are then separated by Capillary Array Electrophoresis (CAE) with automated, *in situ* “finish-line” detection of the fluorescently labeled ssDNA fragments, which provides an ordered sequence of the fragments. These sequence reads are then computer assembled into overlapping or contiguous sequences (termed “contigs”) which resemble the full genomic sequence once fully assembled.^[9]

Sanger methods achieve read lengths of approximately 800bp (typically 500-600bp with non-enriched DNA). The longer read lengths in Sanger methods display significant advantages over other sequencing methods especially in terms of sequencing repetitive regions of the genome. A challenge of short-read sequence data is particularly an issue in sequencing new genomes (*de novo*) and in sequencing highly rearranged genome segments, typically those seen of cancer genomes or in regions of chromosomes that exhibit structural variation.^[10]

Applications of microfluidic sequencing technologies

Other useful applications of DNA sequencing include single nucleotide polymorphism (SNP) detection, single-strand conformation polymorphism (SSCP) hetroduplex analysis, and short tandem repeat (STR) analysis. Resolving DNA fragments according to differences in size and/or conformation is the most critical step in studying these features of the genome.^[9]

Device design

The sequencing chip has a four-layer construction, consisting of three 100-mm-diameter glass wafers (on which device elements are microfabricated) and a polydimethylsiloxane (PDMS) membrane. Reaction chambers and capillary electrophoresis channels are etched between the top two glass wafers, which are thermally bonded. Three-dimensional channel interconnections and microvalves are formed by the PDMS and bottom manifold glass wafer. The device consists of three functional units, each corresponding to the Sanger sequencing steps. The Thermal Cycling (TC) unit is a 250-nanoliter reaction chamber with integrated resistive temperature detector, microvalves, and a surface heater. Movement of reagent between the top all-glass layer and the lower glass-PDMS layer occurs through 500- μ m-diameter via-holes. After thermal-cycling, the reaction mixture undergoes purification in the capture/purification chamber, and then is injected into the capillary electrophoresis (CE) chamber. The CE unit consists of a 30-cm capillary which is folded into a compact switchback pattern via 65- μ m-wide turns.

Sequencing chemistry

- **Thermal cycling**

In the TC reaction chamber, dye-terminator sequencing reagent, template DNA, and primers are loaded into the TC chamber and thermal-cycled for 35 cycles (at 95°C for 12 seconds and at 60°C for 55 seconds).

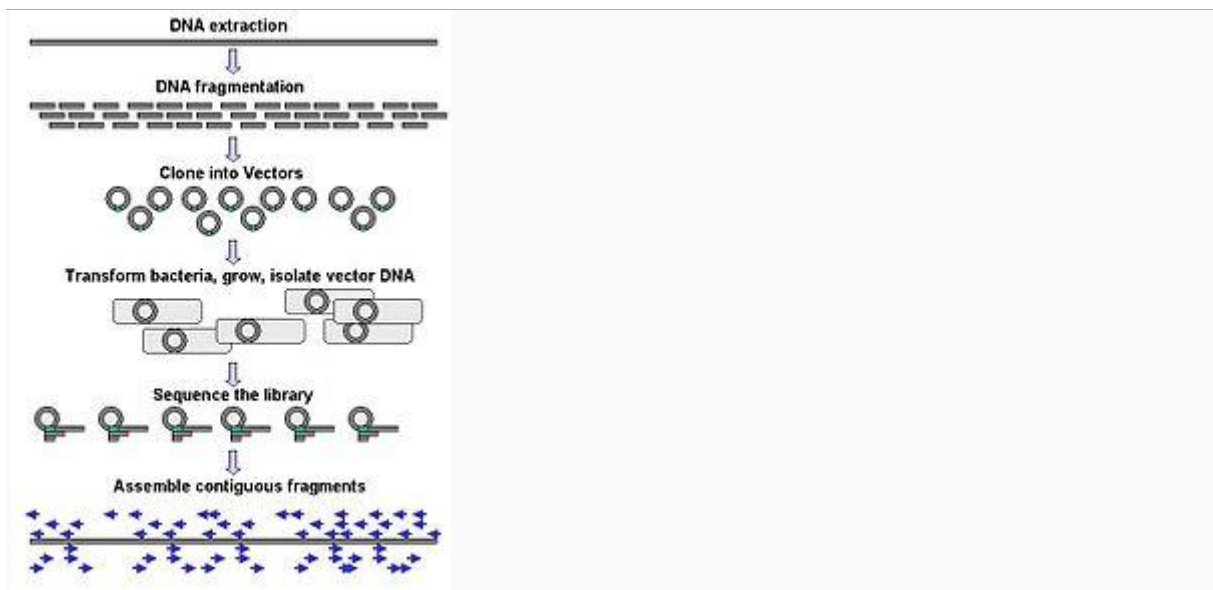
- **Purification**

The charged reaction mixture (containing extension fragments, template DNA, and excess sequencing reagent) is conducted through a capture/purification chamber at 30°C via a 33-Volts/cm electric field applied between capture outlet and inlet ports. The capture gel through which the sample is driven, consists of 40 μ M of oligonucleotide (complementary to the primers) covalently bound to a polyacrylamide matrix. Extension fragments are immobilized by the gel matrix, and excess primer, template, free nucleotides, and salts are eluted through the capture waste port. The capture gel is heated to 67-75°C to release extension fragments.

- **Capillary electrophoresis**

Extension fragments are injected into the CE chamber where they are electrophoresed through a 125-167-V/cm field.

Advanced methods and *de novo* sequencing



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions.(click to expand)

Large-scale sequencing often aims at sequencing very long DNA pieces, such as whole chromosomes, although large-scale sequencing can also be used to generate very large numbers of short sequences, such as found in phage display. For longer targets such as chromosomes, common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented

DNA may then be cloned into a DNA vector and amplified in a bacterial host such as *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence. Studies have shown that adding a size selection step to collect DNA fragments of uniform size can improve sequencing efficiency and accuracy of the genome assembly. In these studies, automated sizing has proven to be more reproducible and precise than manual gel sizing.^{[42][43][44]}

The term "*de novo* sequencing" specifically refers to methods used to determine the sequence of DNA with no previously known sequence. *De novo* translates from Latin as "from the beginning". Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; shotgun methods are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR^[45] isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. A polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods developed by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, later Applied Biosystems, now Life Technologies).^{[46][47][48]}

Shotgun sequencing

In genetics, **shotgun sequencing**, also known as **shotgun cloning**, is a method used for sequencing long DNA strands. It is named by analogy with the rapidly expanding, quasi-random firing pattern of a shotgun.

The chain termination method of DNA sequencing (or "Sanger sequencing" for its developer Frederick Sanger) can only be used for fairly short strands of 100 to 1000 base pairs. Longer sequences are subdivided into smaller fragments that can be sequenced separately, and subsequently they are re-assembled to give the overall sequence. Two principal methods are used for this: primer walking (or "chromosome walking") which progresses through the entire strand piece by piece, and shotgun sequencing which is a faster but more complex process that uses random fragments.

In shotgun sequencing,^{[1][2]} DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain *reads*. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence.^[1]

Shotgun sequencing was one of the precursor technologies that was responsible for enabling full genome sequencing

Whole genome shotgun sequencing

Whole genome shotgun sequencing for small (4000- to 7000-base-pair) genomes was already in use in 1979.^[1] Broader application benefited from pairwise end sequencing, known colloquially as *double-barrel shotgun sequencing*. As sequencing projects began to take on longer and more complicated DNA sequences, multiple groups began to realize that useful information could be obtained by sequencing both ends of a fragment of DNA. Although sequencing both ends of the same fragment and keeping track of the paired data was more cumbersome than sequencing a single end of two distinct fragments, the knowledge that the two sequences were oriented in opposite directions and were about the length of a fragment apart from each other was valuable in reconstructing the sequence of the original target fragment. The first published description of the use of paired ends was in 1990^[4] as part of the sequencing of the human HGPRT locus, although the use of paired ends was limited to closing gaps after the application of a traditional shotgun sequencing approach. The first theoretical description of a pure pairwise end sequencing strategy, assuming fragments of constant length, was in 1991.^[5] At the time, there was community consensus that the optimal fragment length for pairwise end sequencing would be three times the sequence read length. In 1995 Roach et al.^[6] introduced the innovation of using fragments of varying sizes, and demonstrated that a pure pairwise end-sequencing strategy would be possible on large targets. The strategy was subsequently adopted by The Institute for Genomic Research (TIGR) to sequence the genome of the bacterium *Haemophilus influenzae* in 1995,^[7] and then by Celera Genomics to sequence the *Drosophila melanogaster* (fruit fly) genome in 2000,^[8] and subsequently the human genome.

To apply the strategy, a high-molecular-weight DNA strand is sheared into random fragments, size-selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends using the chain termination method yielding two short sequences. Each sequence is called an *end-read* or *read* and two reads from the same clone are referred to as *mate pairs*. Since the chain termination method

usually can only produce reads between 500 and 1000 bases long, in all but the smallest clones, mate pairs will rarely overlap.

The original sequence is reconstructed from the reads using sequence assembly software. First, overlapping reads are collected into longer composite sequences known as *contigs*. Contigs can be linked together into *scaffolds* by following connections between mate pairs. The distance between contigs can be inferred from the mate pair positions if the average fragment length of the library is known and has a narrow window of deviation. Depending on the size of the gap between contigs, different techniques can be used to find the sequence in the gaps. If the gap is small (5-20kb) then the use of PCR to amplify the region is required, followed by sequencing. If the gap is large (>20kb) then the large fragment is cloned in special vectors such as BAC (Bacterial artificial chromosomes) followed by sequencing of the vector.

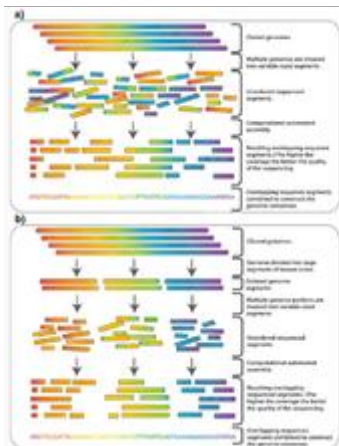
Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than more traditional approaches. Detractors argue that although the technique quickly sequences large regions of DNA, its ability to correctly link these regions is suspect, particularly for genomes with repeating regions. As sequence assembly programs become more sophisticated and computing power becomes cheaper, it may be possible to overcome this limitation.^[citation needed]

Coverage

Coverage (read depth or depth) is the average number of reads representing a given nucleotide in the reconstructed sequence. It can be calculated from the length of the original genome (G), the number of reads (N), and the average read length (L) as $N \times L / G$. For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2x redundancy. This parameter also enables one to estimate other quantities, such as the percentage of the genome covered by reads (sometimes also called coverage). A high coverage in shotgun sequencing is desired because it can overcome errors in base calling and assembly. The subject of DNA sequencing theory addresses the relationships of such quantities.

Sometimes a distinction is made between *sequence coverage* and *physical coverage*. Sequence coverage is the average number of times a base is read (as described above). Physical coverage is the average number of times a base is read or spanned by mate paired reads.

Hierarchical Shotgun sequencing

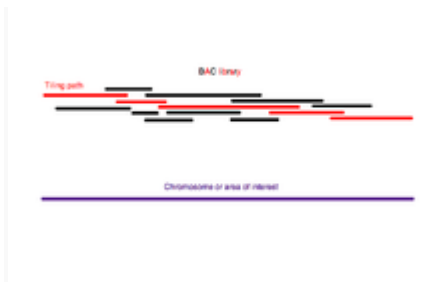


In whole genome shotgun sequencing (top), the entire genome is sheared randomly into small fragments (appropriately sized for sequencing) and then reassembled. In hierarchical shotgun sequencing (bottom), the genome is first broken into larger segments. After the order of these segments is deduced, they are further sheared into fragments appropriately sized for sequencing.

Although shotgun sequencing can in theory be applied to a genome of any size, its direct application to the sequencing of large genomes (for instance, the Human Genome) was limited until the late 1990s, when technological advances made practical the handling of the vast quantities of complex data involved in the process.^[10] Historically, full-genome shotgun sequencing was believed to be limited by both the sheer size of large genomes and by the complexity added by the high percentage of repetitive DNA (greater than 50% for the human genome) present in large genomes.^[11] It was not widely accepted that a full-genome shotgun sequence of a large genome would provide reliable data. For these reasons, other strategies that lowered the computational load of sequence assembly had to be utilized before shotgun sequencing was performed.^[11] In hierarchical sequencing, also known as top-down sequencing, a low-resolution physical map of the genome is made prior to actual sequencing. From this map, a minimal number of fragments that cover the entire chromosome are selected for sequencing.^[12] In this way, the minimum amount of high-throughput sequencing and assembly is required.

The amplified genome is first sheared into larger pieces (50-200kb) and cloned into a bacterial host using BACs or PACs. Because multiple genome copies have been sheared at random, the fragments contained in these clones have different ends, and with enough

coverage (see section above) finding a **scaffold** of BAC contigs that covers the entire genome is theoretically possible. This scaffold is called a **tiling path**.



A BAC contig that covers the entire genomic area of interest makes up the tiling path.

Once a tiling path has been found, the BACs that form this path are sheared at random into smaller fragments and can be sequenced using the shotgun method on a smaller scale.

Although the full sequences of the BAC contigs is not known, their orientations relative to one another are known. There are several methods for deducing this order and selecting the BACs that make up a tiling path. The general strategy involves identifying the positions of the clones relative to one another and then selecting the least number of clones required to form a contiguous scaffold that covers the entire area of interest. The order of the clones is deduced by determining the way in which they overlap.^[13] Overlapping clones can be identified in several ways. A small radioactively or chemically labeled probe containing a sequence-tagged site (STS) can be hybridized onto a microarray upon which the clones are printed.^[13] In this way, all the clones that contain a particular sequence in the genome are identified. The end of one of these clones can then be sequenced to yield a new probe and the process repeated in a method called chromosome walking.

Alternatively, the BAC library can be restriction-digested. Two clones that have several fragment sizes in common are inferred to overlap because they contain multiple similarly spaced restriction sites in common.^[13] This method of genomic mapping is called restriction fingerprinting because it identifies a set of restriction sites contained in each clone. Once the overlap between the clones has been found and their order relative to the genome known, a scaffold of a minimal subset of these contigs that covers the entire genome is shotgun-sequenced.^[12]

Because it involves first creating a low-resolution map of the genome, hierarchical shotgun sequencing is slower than whole-genome shotgun sequencing, but relies less heavily on computer algorithms than whole-genome shotgun sequencing. The process of extensive BAC library creation and tiling path selection, however, make hierarchical shotgun sequencing slow and labor-intensive. Now that the technology is available and the reliability of the data

demonstrated,^[11] and the speed and cost efficiency of whole-genome shotgun sequencing has made it the primary method for genome sequencing.

Shotgun and Next-generation sequencing

The classical shotgun sequencing was based on the Sanger sequencing method: this was the most advanced technique for sequencing genomes from about 1995–2005. The shotgun strategy is still applied today, however using other sequencing technologies, called next-generation sequencing. These technologies produce shorter reads (anywhere from 25–500bp) but many hundreds of thousands or millions of reads in a relatively short time (on the order of a day).^[14] This results in high coverage, but the assembly process is much more computationally intensive. These technologies are vastly superior to Sanger sequencing due to the high volume of data and the relatively short time it takes to sequence a whole genome

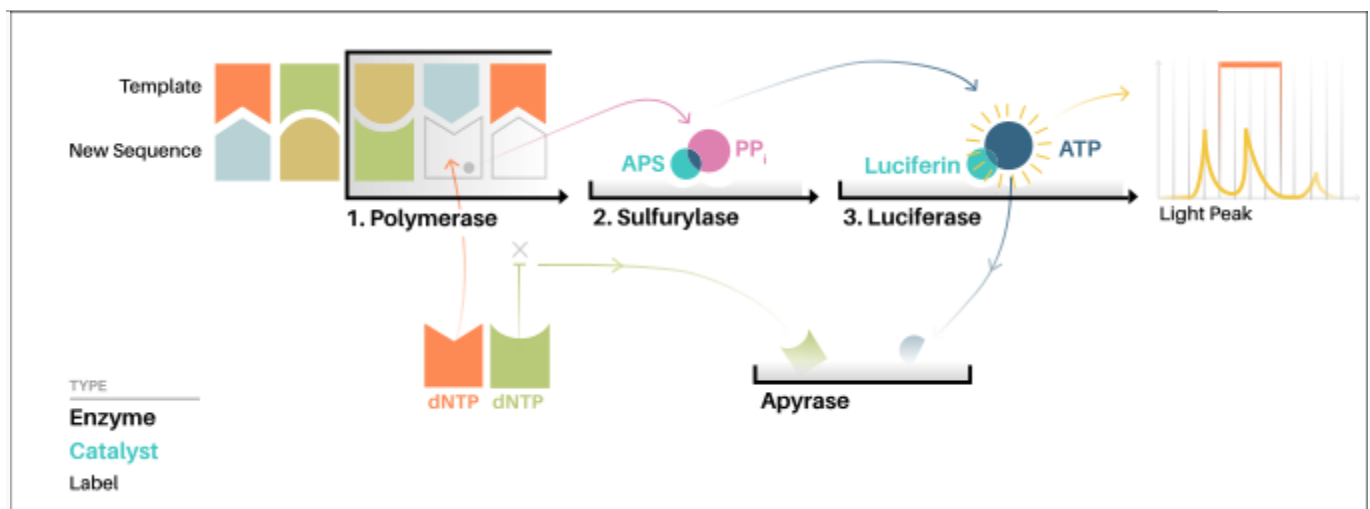
Bridge PCR

Another method for *in vitro* clonal amplification is bridge PCR, in which fragments are amplified upon primers attached to a solid surface^{[35][50][51]} and form "DNA colonies" or "DNA clusters". This method is used in the Illumina Genome Analyzer sequencers. Single-molecule methods, such as that developed by Stephen Quake's laboratory (later commercialized by Helicos) are an exception: they use bright fluorophores and laser excitation to detect base addition events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.^[52]

Pyrosequencing

Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA) based on the "sequencing by synthesis" principle. It differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides.^[1] The technique was developed by Mostafa Ronaghi and Pål Nyrén at the Royal Institute of Technology in Stockholm in 1996.^{[2][3][4]} The desired DNA sequence is able to be determined by light emitted upon incorporation of the next complementary nucleotide by the fact that only one out of four of the possible A/T/C/G nucleotides are added and available at a time so that only one letter can be incorporated on the single stranded template (which is the sequence to be determined). The intensity of the light determines if there are more than one of these "letters" in a row. The previous nucleotide letter (one out of four possible dNTP) is degraded before the next nucleotide letter is added for synthesis: allowing for the possible revealing of the next nucleotide(s) via the resulting intensity of light (if the nucleotide added was the next complementary letter in the sequence). This process is repeated with each of the four letters until the DNA sequence of the single stranded template is determined.

Procedure



The chart shows how pyrosequencing works.

"Sequencing by synthesis" involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The pyrosequencing method is based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time,

and detecting which base was actually added at each step. The template DNA is immobile, and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template.

The single-strand DNA (ssDNA) template is hybridized to a sequencing primer and incubated with the enzymes DNA polymerase, ATP sulfurylase, luciferase and apyrase, and with the substrates adenosine 5' phosphosulfate (APS) and luciferin.

1. The addition of one of the four deoxynucleoside triphosphates (dNTPs) (dATP α S, which is not a substrate for a luciferase, is added instead of dATP to avoid noise) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi).
2. ATP sulfurylase converts PPi to ATP in the presence of adenosine 5' phosphosulfate. This ATP acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a pyrogram.
3. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide.

Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA. As of 2007, pyrosequencing is most commonly used for resequencing or sequencing of genomes for which the sequence of a close relative is already available.

The templates for pyrosequencing can be made both by solid phase template preparation (streptavidin-coated magnetic beads) and enzymatic template preparation (apyrase+exonuclease). So Pyrosequencing can be differentiated into two types, namely Solid Phase Pyrosequencing and Liquid Phase Pyrosequencing.

Commercialization

The company Pyrosequencing AB in Uppsala, Sweden was founded with venture capital provided by HealthCap in order to commercialize machinery and reagents for sequencing short stretches of DNA using the pyrosequencing technique. Pyrosequencing AB was listed on the Stockholm Stock Exchange in 1999. It was renamed to Biotage in 2003. The

pyrosequencing business line was acquired by Qiagen in 2008.^[5] Pyrosequencing technology was further licensed to 454 Life Sciences. 454 developed an array-based pyrosequencing technology which has emerged as a platform for large-scale DNA sequencing. Most notable are the applications for genome sequencing and metagenomics. *GS FLX*, the latest pyrosequencing platform by 454 Life Sciences (now owned by Roche Diagnostics), can generate 400 Mb in a 10-hour run with a single machine. Each run would cost about 5,000-7,000 USD.

Genome databases

A genome comprises all of the genetic material in the chromosomes of a particular organism. Genome databases are an organized collection of information that have resulted from the production or mapping of genome (sequence) or genome product (transcript, protein) information. These databases collect genome sequences, annotate and analyze them, and provide public access. Some add curation of experimental literature to improve computed annotations. These databases may hold many species genomes, or a single model organism genome.

Human Genome Databases, Browsers and Variation Resources

- Database of Genomic Variants
- dbVar Database of Genomic Structural Variation
- ENCODE Project ENCyclopedia Of DNA Elements
- Ensembl Human human genes generated automatically by the Ensembl gene builder
- Entrez Gene searchable database of genes, defined by sequence and/or located in the NCBI Map Viewer
- Genome Reference Consortium Putting sequences into a chromosome context
- GWAS Central centralized compilation of summary level findings from genetic association studies
- HapMap international HapMap Project
- H-Invitational Database an integrated database of human genes and transcripts
- Human Genome Segmental Duplication Database
- Human Structural Variation Database
- 1000 Genomes A Deep Catalog of Human Genetic Variation
- UCSC Human Genome Browser Gateway
- VEGA Human manual annotation of finished genome sequence

Other Vertebrate Genome Databases and Browsers

- AgBase a curated, open-source resource for functional analysis of agricultural plant and animal gene products
- AnolisGenome a community resource site for Anolis genomics and genetic studies
- ARKdb species databases includes: Cat, Chicken, Cow, Deer, Horse, Pig, Salmon, Sheep, Tilapia, Turkey
- BirdBase A Database of Avian Genes and Genomes
- Bovmap mapping the Bovine genome
- Lyons Feline & Comparative Genetics
- Chicken Genome Resources
- The Dog Genome Project
- Ensembl genome databases for vertebrates and other eukaryotic species
- Entrez Gene searchable database of genes, from RefSeq genomes, defined by sequence and/or located in the NCBI Map Viewer
- Fugu the Fugu genomics project
- Horse Genome Project
- Kangaroo Genome Project
- lizardbase a centralized and consolidated informatics resource for lizard research
- MGI Mouse Genome Informatics
- National Animal Genome Research Program
- Pig Genome Coordination Program
- Porcine Genome Sequencing Project
- Pig Genome Resources
- Rabbit Genome Resources
- RGD Rat Genome Database
- Tetraodon Genome Browser
- UCSC Genome Bioinformatics
- VEGA Vertebrate Genome Annotation containing manual annotation of vertebrate finished genome sequence
- Xenbase a Xenopus web resource
- ZFIN Zebrafish Information Network

Non-Vertebrate Genome Databases and Browsers

- ANISEED Ascidian Network for InSitu Expression and Embryological Data
- AspGDA *Aspergillus* Genome Database
- BeetleBase the model organism database for *Tribolium castaneum*

- Cacao Genome Database
- *Caenorhabditis* Genome Sequencing Projects
- *Candida* Genome Database
- ChlamydoDB database for the green alga *Chlamydomonas reinhardtii* and related species
- The Cotton Genome Database
- Daphnia Genome Database
- Dendrome A Forest Tree Genome Database
- dictyBase central resource for Dictyostelid genomics
- EcoGene the Database of *Escherichia coli* Sequence and Function
- Ensembl Genomes
- FlyBase a database of the *Drosophila* genome
- GenProtEC *E.Coli* genome and proteome database
- GOBASE the Organelle Genome Database
- Gramene a resource for comparative grass genomics
- HGD Hymenoptera Genome Database
- IGGI International Glossina Genome Initiative
- PomBase a scientific resource for fission yeast
- SGD *Saccharomyces* Genome Database
- SpBase *Strongylocentrotus purpuratus* Sea Urchin Genome Database
- StellaBase *Nematostella vectensis* Genomics Database
- TAIR The Arabidopsis Information Resource
- VectorBase invertebrate vectors of human pathogens
- WormBase the biology and genome of *C. elegans*

Proteomics Databases

- Proteomics Identifications Database (PRIDE) A public repository for proteomics data, containing protein and peptide identifications and their associated supporting evidence as well as details of post-translational modifications. (European Bioinformatics Institute)
- ProteomeXchange provides a coordinated submission of mass spectrometry proteomics data to the main existing proteomics repositories. It includes datasets such as PRIDE, Tranche, and PeptideAtlas.
- jPOSTrepo Japanese ProteOme STandard repository

- ProteomeScout - A public repository of processed proteomics datasets concerning post-translational modifications, includes quantification across conditions (if applicable). Also includes a graphics exports of protein annotations.
- MitoMiner - A mitochondrial proteomics database integrating large-scale experimental datasets from mass spectrometry and GFP studies for 12 species. (MRC Mitochondrial Biology Unit)
- GelMap - A public database of proteins identified on 2D gels (University of Hanover Proteomics Department)
- OWL - A public non-redundant database for protein search, derived from : SWISS PROT, PIR, GenBank(translation) and NRL-3D
- Proteome-pI pre-computed isoelectric points for >5000 proteomes of model organisms

PPI databases

The primary resources for PPI data are individual scientific publications. Several public databases collect published PPI data and provide researchers access to their curated datasets. These usually reference the original publication and the experimental method that determined every individual interaction. Database designers choose to represent these data in different ways, and the wide spectrum of experimental methods makes it difficult to design a single data model to capture all necessary experimental detail. To overcome this problem, the International Molecular Exchange (IMEx; <http://imex.sourceforge.net/>) consortium was formed. IMEx aims to enable the exchange of data and to avoid the duplication of the curation effort. To that end, an XML-based proteomics standard, termed the proteomics standards initiative - molecular interaction (PSI-MI) has been developed [17]. At the time of writing, however, no data had yet been exchanged, and it was therefore necessary to combine PPI data from all available databases using the authors' own scripts to obtain as comprehensive a network as possible.

Here, the focus is on six databases: the Biological General Repository for Interaction Datasets (BioGRID) [18], the Molecular INTeraction database (MINT) [19], the Biomolecular Interaction Network Database (BIND) [20], the Database of Interacting Proteins (DIP) [21], the IntAct molecular interaction database (IntAct)[22] and the Human Protein Reference Database (HPRD)[23] (see Table 1). These databases report only experimentally verified interactions.

PPI databases

Database	URL	Proteins	Interactions	Publications	Organisms
BioGRID	http://www.thebiogrid.org	23,341	90,972	16,369	10
MINT	http://mint.bio.uniroma2.it/mint	27,306	80,039	3,047	144
BIND	http://bond.unleashedinformatics.com	23,643	43,050	6,364	80
DIP	http://dip.doe-mbi.ucla.edu	21,167	53,431	3,193	134
IntAct	http://www.ebi.ac.uk/intact	37,904	129,559	3,166	131
HPRD	http://www.hprd.org	9,182	36,169	18,777	1

DIP, IntAct and MINT are active members of the IMEx initiative; the curation accuracy of these three databases was assessed recently by Cusick *et al.* [24] HPRD focuses entirely on human proteins, providing not only information on protein interactions, but also a variety of protein-specific information, such as post-translational modifications, disease associations and enzyme-substrate relationships. One of the first interaction databases, BIND, initiated in 2001 by the University of Toronto and the University of British Columbia, is part of the Biomolecular Object Network Databank (BOND) and was subsequently acquired by the company Thomson Reuters.

The following comparison is based on complete sets of binary interactions that were downloaded from the individual databases in May 2008. IntAct and MINT derive binary interactions from protein complexes using the spokes model. No other database provided any information on which model is applied. Only 'physical interactions' are considered here, although most databases also provide 'genetic interactions' -- that is, two non-essential genes that lead to a non-viable phenotype if they are knocked out simultaneously. Furthermore, interactions were only accepted if a publication identifier was provided along with the interacting proteins.

Currently, the most comprehensive database in terms of individual interactions is IntAct, with almost 130,000 unique interactions from up to 131 different organisms. Despite these large numbers, it cites only about 3,000 different publications. Whereas IntAct seems to be concentrating on high-throughput studies, HPRD also takes into account small-scale publications. Although being restricted to human proteins, it reports over 36,000 unique interactions from more than 18,000 publications. Only BioGRID cites a similar number of publications (16,369); it is also the second largest database in terms of the number of unique interactions. It should be noted that the databases examine publications in different depth, and that higher numbers of publications do not necessarily involve a higher curation effort.

The majority of known protein interactions account for proteins from *Saccharomyces cerevisiae* and *Homo sapiens*. Individual high-throughput interaction screens were carried out for some other organisms; these high-throughput studies usually account for the majority of all known interactions in the corresponding organism. By contrast, known protein interactions for *S. cerevisiae* and *H. sapiens* are dispersed over numerous publications. For this reason, the number of interactions for humans and yeast can vary considerably between different databases, depending on their coverage of the literature.



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY**

UNIT – II - Fundamentals of Genomics and Proteomics– SBI1309

Genome Mapping

Assigning/locating of a specific gene to particular region of a chromosome and determining the location of and relative distances between genes on the chromosome.

The convention is to divide genome mapping methods into two categories.

- Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome. Genetic techniques include cross-breeding experiments or, in the case of humans, the examination of family histories (pedigrees). Genetic mapping is described in Section 5.2.
- Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features, including genes. Physical mapping is described in Section 5.3.


5.2. Genetic Mapping

As with any type of map, a genetic map must show the positions of distinctive features. In a geographic map these markers are recognizable components of the landscape, such as rivers, roads and buildings. What markers can we use in a genetic landscape?

5.2.1. Genes were the first markers to be used

The first genetic maps, constructed in the early decades of the 20th century for organisms such as the fruit fly, used genes as markers. This was many years before it was understood that genes are segments of DNA molecules. Instead, genes were looked upon as abstract entities responsible for the transmission of heritable characteristics from parent to offspring. To be useful in genetic analysis, a heritable characteristic has to exist in at least two alternative forms or phenotypes, an example being tall or short stems in the pea plants originally studied by Mendel. Each phenotype is specified by a different allele of the corresponding gene. To begin with, the only genes that could be studied were those specifying phenotypes that were distinguishable by visual examination. So, for example, the first fruit-fly maps showed the positions of genes for body color, eye color, wing shape and suchlike, all of these phenotypes being visible simply by looking at the flies with a low-power microscope or the naked eye. This approach was fine in the early days but geneticists soon realized that there were only a limited number of visual phenotypes whose inheritance could be studied, and in many cases their analysis was complicated because a single phenotype could be affected by more than one gene. For example, by 1922 over 50 genes had been mapped onto the four fruit-fly chromosomes, but nine of these

were for eye color; in later research, geneticists studying fruit flies had to learn to distinguish between fly eyes that were colored red, light red, vermilion, garnet, carnation, cinnabar, ruby, sepia, scarlet, pink, cardinal, claret, purple or brown. To make gene maps more comprehensive it would be necessary to find characteristics that were more distinctive and less complex than visual ones.

The answer was to use biochemistry to distinguish phenotypes. This has been particularly important with two types of organisms - microbes and humans. Microbes, such as bacteria and yeast, have very few visual characteristics so gene mapping with these organisms has to rely on biochemical phenotypes such as those listed in [Table 5.1](#) . With humans it is possible to use visual characteristics, but since the 1920s studies of human genetic variation have been based largely on biochemical phenotypes that can be scored by blood typing. These phenotypes include not only the standard blood groups such as the ABO series ( Yamamoto *et al.*, 1990), but also variants of blood serum proteins and of immunological proteins such as the human leukocyte antigens (the HLA system). A big advantage of these markers is that many of the relevant genes have multiple alleles. For example, the gene called *HLA-DRB1* has at least 290 alleles and *HLA-B* has over 400. This is relevant because of the way in which gene mapping is carried out with humans ([Section 5.2.4](#)). Rather than setting up many breeding experiments, which is the procedure with experimental organisms such as fruit flies or mice, data on inheritance of human genes have to be gleaned by examining the phenotypes displayed by members of a single family. If all the family members have the same allele for the gene being studied then no useful information can be obtained. It is therefore necessary for the relevant marriages to have occurred, by chance, between individuals with different alleles. This is much more likely if the gene being studied has 290 rather than two alleles.

5.2.2. DNA markers for genetic mapping

Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed. This would be true even if every gene could be mapped because, as we saw in [Chapter 2](#), in most eukaryotic genomes the genes are widely spaced out with large gaps between them (see [Figure 2.2](#)). The problem is made worse by the fact that only a fraction of the total number of genes exist in allelic forms that can be distinguished conveniently. Gene maps are therefore not very comprehensive. We need other types of marker.

Mapped features that are not genes are called DNA markers. As with gene markers, a DNA marker must have at least two alleles to be useful. There are three types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs).

Restriction fragment length polymorphisms (RFLPs)

RFLPs were the first type of DNA marker to be studied. Recall that restriction enzymes cut DNA molecules at specific recognition sequences (Section 4.1.2). This sequence specificity means that treatment of a DNA molecule with a restriction enzyme should always produce the same set of fragments. This is not always the case with genomic DNA molecules because some restriction sites are polymorphic, existing as two alleles, one allele displaying the correct sequence for the restriction site and therefore being cut when the DNA is treated with the enzyme, and the second allele having a sequence alteration so the restriction site is no longer recognized. The result of the sequence alteration is that the two adjacent restriction fragments remain linked together after treatment with the enzyme, leading to a length polymorphism (Figure 5.4). This is an RFLP and its position on a genome map can be worked out by following the inheritance of its alleles, just as is done when genes are used as markers. There are thought to be about 10^5 RFLPs in the human genome, but of course for each RFLP there can only be two alleles (with and without the site). The value of RFLPs in human gene mapping is therefore limited by the high possibility that the RFLP being studied shows no variability among the members of an interesting family.

In order to score an RFLP, it is necessary to determine the size of just one or two individual restriction fragments against a background of many irrelevant fragments. This is not a trivial problem: an enzyme such as *EcoRI*, with a 6-bp recognition sequence, should cut approximately once every $4^6 = 4096$ bp and so would give almost 800 000 fragments when used with human DNA. After separation by agarose gel electrophoresis (see Technical Note 2.1), these 800 000 fragments produce a smear and the RFLP cannot be distinguished. Southern hybridization, using a probe that spans the polymorphic restriction site, provides one way of visualizing the RFLP (Figure 5.5A), but nowadays PCR is more frequently used. The primers for the PCR are designed so that they anneal either side of the polymorphic site, and the RFLP is typed by treating the amplified fragment with the restriction enzyme and then running a sample in an agarose gel (Figure 5.5B).


Simple sequence length polymorphisms (SSLPs)

SSLPs are arrays of repeat sequences that display length variations, different alleles containing different numbers of repeat units (*Figure 5.6A*). Unlike RFLPs, SSLPs can be multi-allelic as each SSLP can have a number of different length variants. There are two types of SSLP, both of which were described in *Section 2.4.1*:


- Minisatellites, also known as variable number of tandem repeats (VNTRs), in which the repeat unit is up to 25 bp in length;
- Microsatellites or **simple tandem repeats (STRs)**, whose repeats are shorter, usually dinucleotide or tetranucleotide units.




Microsatellites are more popular than minisatellites as DNA markers, for two reasons. First, minisatellites are not spread evenly around the genome but tend to be found more frequently in the telomeric regions at the ends of chromosomes. In geographic terms, this is equivalent to trying to use a map of lighthouses to find one's way around the middle of an island. Microsatellites are more conveniently spaced throughout the genome. Second, the quickest way to type a length polymorphism is by PCR (*Figure 5.6B*), but PCR typing is much quicker and more accurate with sequences less than 300 bp in length. Most minisatellite alleles are longer than this because the repeat units are relatively large and there tend to be many of them in a single array, so PCR products of several kb are needed to type them. Typical microsatellites consist of 10–30 copies of a repeat that is usually no longer than 4 bp in length, and so are much more amenable to analysis by PCR. There are 6.5×10^5 microsatellites in the human genome (see *Table 1.3*).

Single nucleotide polymorphisms (SNPs)

These are positions in a genome where some individuals have one nucleotide (e.g. a G) and others have a different nucleotide (e.g. a C) (*Figure 5.7*). There are vast numbers of SNPs in every genome, some of which also give rise to RFLPs, but many of which do not because the sequence in which they lie is not recognized by any restriction enzyme. In the human genome there are at least 1.42 million SNPs, only 100 000 of which result in an RFLP ( *SNP Group, 2001*).

Although each SNP could, potentially, have four alleles (because there are four nucleotides), most exist in just two forms, so these markers suffer from the same drawback as RFLPs with regard to human genetic mapping: there is a high possibility that a SNP does not display any variability in the family that is being studied. The advantages of SNPs are their abundant

numbers and the fact that they can be typed by methods that do not involve gel electrophoresis. This is important because gel electrophoresis has proved difficult to automate so any detection method that uses it will be relatively slow and labor-intensive. SNP detection is more rapid because it is based on oligonucleotide hybridization analysis. An oligonucleotide is a short single-stranded DNA molecule, usually less than 50 nucleotides in length, that is synthesized in the test tube. If the conditions are just right, then an oligonucleotide will hybridize with another DNA molecule only if the oligonucleotide forms a completely base-paired structure with the second molecule. If there is a single mismatch - a single position within the oligonucleotide that does not form a base pair - then hybridization does not occur (*Figure 5.8*). Oligonucleotide hybridization can therefore discriminate between the two alleles of an SNP. Various screening strategies have been devised ( Mir and Southern, 2000), including DNA chip technology (Technical Note 5.1) and **solution hybridization techniques**.

- A DNA chip is a wafer of glass or silicon, 2.0 cm^2 or less in area, carrying many different oligonucleotides in a high-density array. The DNA to be tested is labeled with a fluorescent marker and pipetted onto the surface of the chip. Hybridization is detected by examining the chip with a fluorescence microscope, the positions at which the fluorescent signal is emitted indicating which oligonucleotides have hybridized with the test DNA. Many SNPs can therefore be scored in a single experiment ( Wang *et al.*, 1998;  Gerhold *et al.*, 1999).
- **Solution hybridization techniques** are carried out in the wells of a microtiter tray, each well containing a different oligonucleotide, and use a detection system that can discriminate between unhybridized single-stranded DNA and the double-stranded product that results when an oligonucleotide hybridizes to the test DNA. Several systems have been developed, one of which makes use of a pair of labels comprising a fluorescent dye and a compound that quenches the fluorescent signal when brought into close proximity with the dye. The dye is attached to one end of an oligonucleotide and the quenching compound to the other end. Normally there is no fluorescence because the oligonucleotide is designed in such a way that the two ends base-pair to one another, placing the quencher next to the dye (*Figure 5.9*). Hybridization between oligonucleotide and test DNA disrupts this base pairing, moving the quencher away from the dye and enabling the fluorescent signal to be generated ( Tyagi *et al.*, 1998).

A **single-nucleotide polymorphism (SNP)**, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two *alleles* : C and T. Almost all common SNPs have only two alleles.

Within a population, SNPs can be assigned a minor allele frequency — the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for single-nucleotide polymorphisms. There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another.

Types of SNPs

- Non-coding region
- Coding region
 - Synonymous
 - Nonsynonymous
 - Missense
 - Nonsense

Single nucleotides may be changed (substitution), removed (deletions) or added (insertion) to a polynucleotide sequence. Ins/del SNP may shift translational frame.

Single nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. A SNP in which both forms lead to the same polypeptide sequence is termed *synonymous* (sometimes called a silent mutation) — if a different polypeptide sequence is produced they are *nonsynonymous*. A nonsynonymous change may either be missense or nonsense, where a missense change results in a different amino acid, while a nonsense change results in a premature stop codon. SNPs that are not in protein-coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA.

Use and importance of SNPs

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine. However, their greatest importance in biomedical research is for comparing regions of the genome between cohorts (such as with matched cohorts with and without a disease).

The study of single-nucleotide polymorphisms is also important in crop and livestock breeding programs

Examples

- rs6311 and rs6313 are SNPs in the HTR2A gene on human chromosome 13.
- A SNP in the *F5* gene causes a hypercoagulability disorder with the variant Factor V Leiden.
- rs3091244 is an example of a triallelic SNP in the CRP gene on human chromosome 1.^[6]
- TAS2R38 codes for PTC tasting ability, and contains 6 annotated SNPs.^[citation needed]

Databases

As there are for genes, there are also bioinformatics databases for SNPs. *dbSNP* is a SNP database from National Center for Biotechnology Information (NCBI). *SNPedia* is a wiki-style database from a hybrid organization. The *OMIM* database describes the association between polymorphisms and, e.g., diseases in text form, while HGVbase - (Human Genome Variation Database) - A human gene-based polymorphism database. Records in dbSNP are cross-annotated within other internal information resources such as PubMed, genome project sequences, GenBank records, the Entrez Gene database, and the dbSTS database of sequence tagged sites. Users may query dbSNP directly or start a search in any part of the NCBI discovery space to construct a set of dbSNP records that satisfy their search conditions. Records are also integrated with external information resources through hypertext URLs that dbSNP users can follow to explore the detailed information that is beyond the scope of dbSNP curation.

Nomenclature

The nomenclature for SNPs can be confusing: several variations can exist for an individual SNP and consensus has not yet been achieved. One approach is to write SNPs with a prefix, period and greater than sign showing the wild-type and altered nucleotide or amino acid; for example, c.76A>T.

SNP genotyping

Genotyping provides a measurement of the genetic variation between members of a species. Single nucleotide polymorphisms (SNP) are the most common type of genetic variation. A SNP is a single base pair mutation at a specific locus, usually consisting of two alleles (where the rare allele frequency is $\geq 1\%$). SNPs are often found to be the etiology of many human diseases and are becoming of particular interest in pharmacogenetics. Because SNPs are evolutionarily conserved, they have been proposed as markers for use in quantitative trait loci (QTL) analysis and in association studies in place of microsatellites. The use of SNPs is being extended in the HapMap project, which is attempting to provide the minimal set of SNPs needed to genotype the human genome. SNPs can also provide a genetic fingerprint for use in identity testing (Rapley & Harbron 2004).

The increase in interest in SNPs has been reflected by the furious development of a diverse range of **SNP genotyping** methods. This article provides an overview of the major strategies for interrogating SNPs.

Hybridization-based methods

Several applications have been developed that interrogate SNPs by hybridizing complementary DNA probes to the SNP site. The challenge of this approach is reducing cross-hybridization between the allele-specific probes. This challenge is generally overcome by manipulating the hybridization stringency conditions (Rapley & Harbron 2004).

Dynamic allele-specific hybridization

Dynamic allele-specific hybridization (DASH) genotyping takes advantage of the differences in the melting temperature in DNA that results from the instability of mismatched base pairs. The process can be vastly automated and encompasses a few simple principles.

In the first step, a genomic segment is amplified and attached to a bead through a PCR reaction with a biotinylated primer. In the second step, the amplified product is attached to a streptavidin

column and washed with NaOH to remove the unbiotinylated strand. An allele specific oligonucleotide is then added in the presence of a molecule that fluoresces when bound to double-stranded DNA. The intensity is then measured as temperature is increased until the T_m can be determined. A SNP will result in a lower than expected T_m (Howell et al. 1999).

Because DASH genotyping is measuring a quantifiable change in T_m , it is capable of measuring all types of mutations, not just SNPs. Other benefits of DASH include its ability to work with label free probes and its simple design and performance conditions.

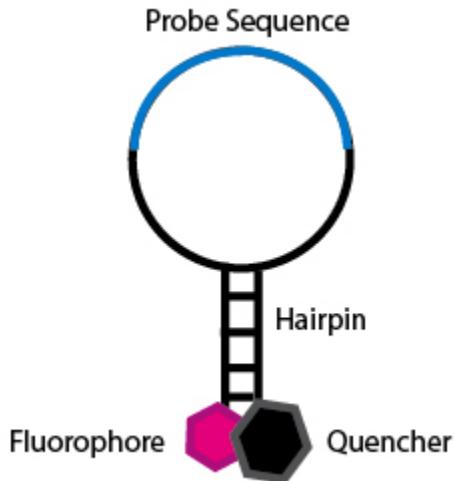
Molecular beacons

Molecular Beacon Probes

Molecular beacons are oligonucleotide hybridization probes that can report the presence of specific nucleic acids in homogenous solutions. The terms more often used is **molecular beacon probes**. Molecular beacons are hairpin shaped molecules with an internally quenched fluorophore whose fluorescence is restored when they bind to a target nucleic acid sequence. This is a novel nonradioactive method for detecting specific sequences of nucleic acids. They are useful in situations where it is either not possible or desirable to isolate the probe-target hybrids from an excess of the hybridization probes.

A typical molecular beacon probe is 25 nucleotides long. The middle 15 nucleotides are complementary to the target DNA and do not base pair with one another, and the five nucleotides at each end are complementary to each other and not to the target DNA. A typical molecular Beacon Structure can be divided in 4 parts :

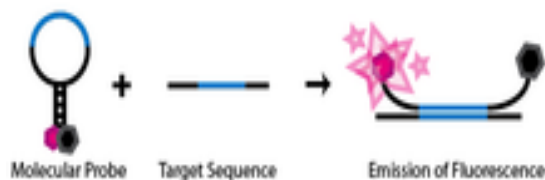
- **Loop:** This is the 18-30 base pair region of the molecular beacon which is complementary to the target sequence.
- **Stem:** The beacon stem sequence lies on both the ends of the loop. It is typically 5-7 bp long at the sequences at both the ends are complementary to each other.
- **5' fluorophore:** Towards the 3' end of the molecular beacon, is attached a dye that fluoresces in presence of a complementary target.
- **3' quencher (non fluorescent):** The quencher dye is covalently attached to the 3' end of the molecular beacon and when the beacon is in closed loop shape, prevents the fluorophore from emitting light.



SNP detection through Molecular beacons makes use of a specifically engineered single-stranded oligonucleotide probe. The oligonucleotide is designed such that there are complementary regions at each end and a probe sequence located in between. This design allows the probe to take on a hairpin, or stem-loop, structure in its natural, isolated state. Attached to one end of the probe is a fluorophore and to the other end a fluorescence quencher. Because of the stem-loop structure of the probe, the fluorophore is in close proximity to the quencher, thus preventing the molecule from emitting any fluorescence. The molecule is also engineered such that only the probe sequence is complementary to the genomic DNA that will be used in the assay (Abravaya et al. 2003).

If the probe sequence of the molecular beacon encounters its target genomic DNA during the assay, it will anneal and hybridize. Because of the length of the probe sequence, the hairpin segment of the probe will denature in favour of forming a longer, more stable probe-target hybrid. This conformational change permits the fluorophore and quencher to be free of their tight proximity due to the hairpin association, allowing the molecule to fluoresce.

If on the other hand, the probe sequence encounters a target sequence with as little as one non-complementary nucleotide, the molecular beacon will preferentially stay in its natural hairpin state and no fluorescence will be observed, as the fluorophore remains quenched.





The unique design of these molecular beacons allows for a simple diagnostic assay to identify SNPs at a given location. If a molecular beacon is designed to match a wild-type allele and another to match a mutant of the allele, the two can be used to identify the genotype of an individual. If only the first probe's fluorophore wavelength is detected during the assay then the individual is homozygous to the wild type. If only the second probe's wavelength is detected then the individual is homozygous to the mutant allele. Finally, if both wavelengths are detected, then both molecular beacons must be hybridizing to their complements and thus the individual must contain both alleles and be heterozygous.

SNP microarrays

In high density oligonucleotide SNP arrays, hundreds of thousands of probes are arrayed on a small chip, allowing for many SNPs to be interrogated simultaneously (Rapley & Harbron 2004). Because SNP alleles only differ in one nucleotide and because it is difficult to achieve optimal hybridization conditions for all probes on the array, the target DNA has the potential to hybridize to mismatched probes. This is addressed somewhat by using several redundant probes to interrogate each SNP. Probes are designed to have the SNP site in several different locations as well as containing mismatches to the SNP allele. By comparing the differential amount of hybridization of the target DNA to each of these redundant probes, it is possible to determine specific homozygous and heterozygous alleles (Rapley & Harbron 2004). Although oligonucleotide microarrays have a comparatively lower specificity and sensitivity, the scale of SNPs that can be interrogated is a major benefit. The Affymetrix Human SNP 5.0 GeneChip performs a genome-wide assay that can genotype over 500,000 human SNPs (Affymetrix 2007).

Enzyme-based methods

A broad range of enzymes including DNA ligase, DNA polymerase and nucleases have been employed to generate high-fidelity SNP genotyping methods.

Restriction fragment length polymorphism

Restriction fragment length polymorphism (RFLP) is considered to be the simplest and earliest method to detect SNPs. SNP-RFLP makes use of the many different restriction endonucleases

and their high affinity to unique and specific restriction sites. By performing a digestion on a genomic sample and determining fragment lengths through a gel assay it is possible to ascertain whether or not the enzymes cut the expected restriction sites. A failure to cut the genomic sample results in an identifiably larger than expected fragment implying that there is a mutation at the point of the restriction site which is rendering it protected from nuclease activity.

Unfortunately, the combined factors of the high complexity of most eukaryotic genomes, the requirement for specific endonucleases, the fact that the exact mutation cannot be necessarily be resolved in a single experiment, and the slow nature of gel assays make RFLP a poor choice for high throughput analysis.

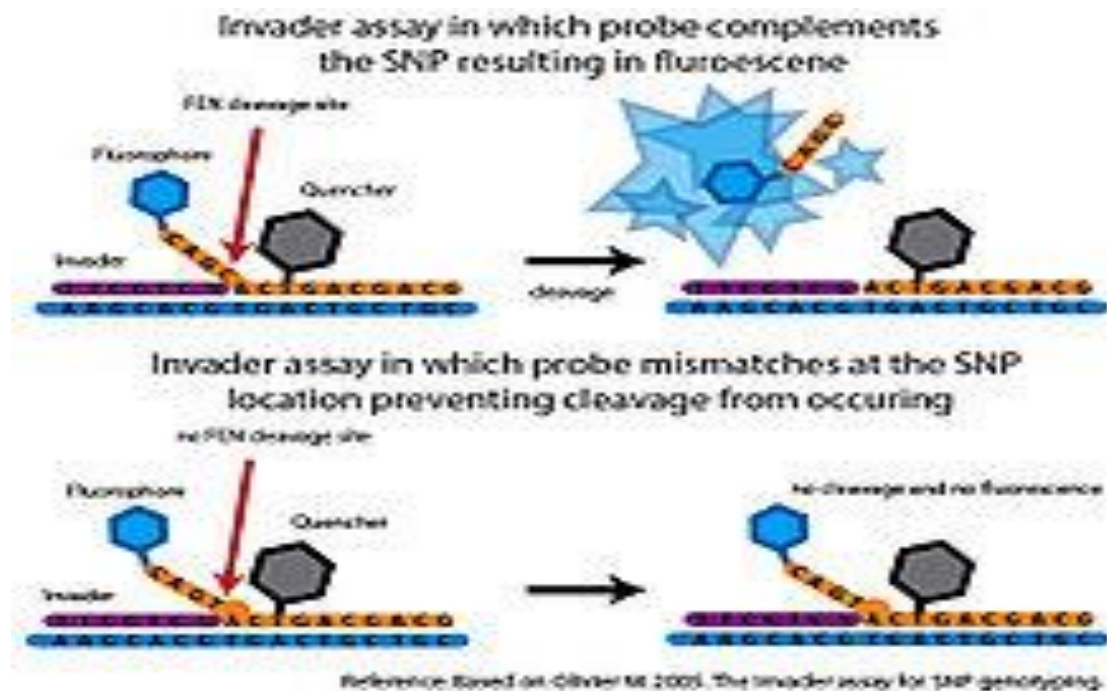
PCR-based methods

Tetra-primer ARMS-PCR employs two pairs of primers to amplify two alleles in one PCR reaction. The primers are designed such that the two primer pairs overlap at a SNP location but each match perfectly to only one of the possible SNPs. As a result, if a given allele is present in the PCR reaction, the primer pair specific to that allele will produce product but not to the alternative allele with a different SNP. The two primer pairs are also designed such that their PCR products are of a significantly different length allowing for easily distinguishable bands by gel electrophoresis.

In examining the results, if a genomic sample is homozygous, then the PCR products that result will be from the primer which matches the SNP location to the outer, opposite strand primer as well from the two opposite, outer primers. If the genomic sample is heterozygous, then products will result from the primer of each allele to their respective outer primer counterparts as well as from the two opposite, outer primers.

The difficulty in designing multiple pairs of primers for a single PCR reaction is vastly outweighed by the simplicity and speed at which samples can be examined.

Flap endonuclease



Flap endonuclease (FEN) is an endonuclease that catalyzes structure-specific cleavage. This cleavage is highly sensitive to mismatches and can be used to interrogate SNPs with a high degree of specificity (Olivier 2005).

In the basic **Invader** assay, a FEN called cleavase is combined with two specific oligonucleotide probes, that together with the target DNA, can form a tripartite structure recognized by cleavase (Olivier 2005). The first probe, called the **Invader** oligonucleotide is complementary to the 3' end of the target DNA. The last base of the **Invader** oligonucleotide is a non-matching base that overlaps the SNP nucleotide in the target DNA. The second probe is an allele-specific probe which is complementary to the 5' end of the target DNA, but also extends past the 3' side of the SNP nucleotide. The allele-specific probe will contain a base complementary to the SNP nucleotide. If the target DNA contains the desired allele, the Invader and allele-specific probes will bind to the target DNA forming the tripartite structure. This structure is recognized by cleavase, which will cleave and release the 3' end of the allele-specific probe. If the SNP nucleotide in the target DNA is not complementary allele-specific probe, the correct tripartite structure is not formed and no cleavage occurs. The **Invader** assay is usually coupled with fluorescence resonance energy transfer (FRET) system to detect the cleavage event. In this setup, a quencher molecule is attached to the 3' end and a fluorophore is attached to the 5' end of the

allele-specific probe. If cleavage occurs, the fluorophore will be separated from the quencher molecule generating a detectable signal (Olivier 2005).

When cleavage by FEN generates a detectable fluorescent signal, the signal is measured using flow-cytometry. The sensitivity of flow-cytometry, eliminates the need for PCR amplification of the target DNA (Rao et al. 2003). These high-throughput platforms have not progressed beyond the proof-of-principle stage and so far the **Invader** system has not been used in any large scale SNP genotyping projects (Olivier 2005).

Primer extension

Primer extension is a two step process that first involves the hybridization of a probe to the bases immediately upstream of the SNP nucleotide followed by a ‘mini-sequencing’ reaction, in which DNA polymerase extends the hybridized primer by adding a base that is complementary to the SNP nucleotide. This incorporated base is detected and determines the SNP allele (Goelet et al. 1999; Syvanen 2001). Because, primer extension is based on the highly accurate DNA polymerase enzyme, the method is generally very reliable. Primer extension is able to genotype most SNPs under very similar reaction conditions making it also highly flexible. The primer extension method is used in a number of assay formats. These formats use a wide range of detection techniques that include MALDI-TOF Mass spectrometry (see Sequenom) and ELISA-like methods (Rapley & Harbron 2004).

Generally, there are two main approaches which use the incorporation of either fluorescently labeled dideoxynucleotides (ddNTP) or fluorescently labeled deoxynucleotides (dNTP). With ddNTPs, probes hybridize to the target DNA immediately upstream of SNP nucleotide, and a single, ddNTP complementary to the SNP allele is added to the 3’ end of the probe (the missing 3’-hydroxyl in dideoxynucleotide prevents further nucleotides from being added). Each ddNTP is labeled with a different fluorescent signal allowing for the detection of all four alleles in the same reaction. With dNTPs, allele-specific probes have 3’ bases which are complementary to each of the SNP alleles being interrogated. If the target DNA contains an allele complementary to the probe's 3’ base, the target DNA will completely hybridize to the probe, allowing DNA polymerase to extend from the 3’ end of the probe. This is detected by the incorporation of the fluorescently labeled dNTPs onto the end of the probe. If the target DNA does not contain an allele complementary to the probe's 3’ base, the target DNA will produce a

mismatch at the 3' end of the probe and DNA polymerase will not be able to extend from the 3' end of the probe. The benefit of the second approach is that several labeled dNTPs may get incorporated into the growing strand, allowing for increased signal. However, DNA polymerase in some rare cases, can extend from mismatched 3' probes giving a false positive result (Rapley & Harbron 2004).

5'- nuclease

Taq DNA polymerase's 5'-nuclease activity is used in the **Taqman** assay for SNP genotyping. The **Taqman** assay is performed concurrently with a PCR reaction and the results can be read in real-time as the PCR reaction proceeds. The assay requires forward and reverse PCR primers that will amplify a region that includes the SNP polymorphic site. Allele discrimination is achieved using FRET combined with one or two allele-specific probes that hybridize to the SNP polymorphic site. The probes will have a fluorophore linked to their 5' end and a quencher molecule linked to their 3' end. While the probe is intact, the quencher will remain in close proximity to the fluorophore, eliminating the fluorophore's signal. During the PCR amplification step, if the allele-specific probe is perfectly complementary to the SNP allele, it will bind to the target DNA strand and then get degraded by 5'-nuclease activity of the Taq polymerase as it extends the DNA from the PCR primers. The degradation of the probe results in the separation of the fluorophore from the quencher molecule, generating a detectable signal. If the allele-specific probe is not perfectly complementary, it will have lower melting temperature and not bind as efficiently. This prevents the nuclease from acting on the probe (McGuigan & Ralston 2002).

Since the **Taqman** assay is based on PCR, it is relatively simple to implement. The **Taqman** assay can be multiplexed by combining the detection of up to seven SNPs in one reaction. However, since each SNP requires a distinct probe, the **Taqman** assay is limited by the how close the SNPs can be situated. The scale of the assay can be drastically increased by performing many simultaneous reactions in microtitre plates. Generally, **Taqman** is limited to applications that involve interrogating a small number of SNPs since optimal probes and reaction conditions must be designed for each SNP (Syvanen 2001).

Oligonucleotide ligase assay

DNA ligase catalyzes the ligation of the 3' end of a DNA fragment to the 5' end of a directly adjacent DNA fragment. This mechanism can be used to interrogate a SNP by hybridizing two probes directly over the SNP polymorphic site, whereby ligation can occur if the probes are identical to the target DNA. In the oligonucleotide ligase assay, two probes are designed; an allele-specific probe which hybridizes to the target DNA so that its 3' base is situated directly over the SNP nucleotide and a second probe that hybridizes the template upstream (downstream in the complementary strand) of the SNP polymorphic site providing a 5' end for the ligation reaction. If the allele-specific probe matches the target DNA, it will fully hybridize to the target DNA and ligation can occur. Ligation does not generally occur in the presence of a mismatched 3' base. Ligated or unligated products can be detected by gel electrophoresis, MALDI-TOF mass spectrometry or by capillary electrophoresis for large-scale applications (Rapley & Harbron 2004).

SNP array

In molecular biology and bioinformatics, a **SNP array** is a type of DNA microarray which is used to detect polymorphisms within a population. A single nucleotide polymorphism (SNP), a variation at a single site in DNA, is the most frequent type of variation in the genome. For example, there are around 10 million SNPs that have been identified in the human genome^[1]. As SNPs are highly conserved throughout evolution and within a population, the map of SNPs serves as an excellent genotypic marker for research.

Principles

The basic principles of SNP array are the same as the DNA microarray. These are the convergence of DNA hybridization, fluorescence microscopy, and solid surface DNA capture. The three mandatory components of the SNP arrays are:

1. The array that contains immobilized nucleic acid sequences or target;
2. One or more labeled Allele specific oligonucleotide (ASO) probes;
3. A detection system that records and interprets the hybridization signal.

To achieve relative concentration independence and minimal cross-hybridization, raw sequences and SNPs of multiple databases are scanned to design the probes. Each SNP on the array is interrogated with different probes. Depending on the purpose of experiments, the amount of SNPs present on an array is considered.

Applications

An SNP array is a useful tool to study the whole genome. The most important application of SNP array is in determining disease susceptibility and consequently, in pharmacogenomics by measuring the efficacy of drug therapies specifically for the individual. As each individual has many single nucleotide polymorphisms that together create a unique DNA sequence, SNP-based genetic linkage analysis could be performed to map disease loci, and hence determine disease susceptibility genes for an individual. The combination of SNP maps and high density SNP array allows the use of SNPs as the markers for Mendelian diseases with complex traits efficiently. For example, whole-genome genetic linkage analysis shows significant linkage for many diseases such as rheumatoid arthritis, prostate cancer, and neonatal diabetes. As a result, drugs can be personally designed to efficiently act on a group of individuals who share a common allele - or even a single individual. A SNP array can also be used to generate a virtual karyotype using specialized software to determine the copy number of each SNP on the array and then align the SNPs in chromosomal order.

In addition, SNP array can be used for studying the Loss of heterozygosity (LOH). LOH is a form of allelic imbalance that can result from the complete loss of an allele or from an increase in copy number of one allele relative to the other. While other chip-based methods (e.g. Comparative genomic hybridization) can detect only genomic gains or deletions, SNP array has the additional advantage of detecting copy number neutral LOH due to uniparental disomy (UPD). In UPD, one allele or whole chromosome from one parent are missing leading to reduplication of the other parental allele (uni-parental = from one parent, disomy = duplicated). In a disease setting this occurrence may be pathologic when the wildtype allele (e.g. from the mother) is missing and instead two copies of the mutant allele (e.g. from the father) are present. Using high density SNP array to detect LOH allows identification of pattern of allelic imbalance with potential prognostic and diagnostic utilities. This usage of SNP array has a huge potential in cancer diagnostics as LOH is a prominent characteristic of most human cancers. Recent studies based on the SNP array technology have shown that not only solid tumors (e.g. gastric cancer, liver cancer etc) but also hematologic malignancies (ALL, MDS, CML etc) have a high rate of LOH due to genomic deletions or UPD and genomic gains. The results of these studies may help to gain insights into mechanisms of these diseases and to create targeted drugs.

5.2.3. Linkage analysis is the basis of genetic mapping

Now that we have assembled a set of markers with which to construct a genetic map we can move on to look at the mapping techniques themselves. These techniques are all based on genetic linkage, which in turn derives from the seminal discoveries in genetics made in the mid 19th century by Gregor Mendel.

The principles of inheritance and the discovery of linkage

Genetic mapping is based on the principles of inheritance as first described by Gregor Mendel in 1865 (Orel, 1995). From the results of his breeding experiments with peas, Mendel concluded that each pea plant possesses two alleles for each gene, but displays only one phenotype. This is easy to understand if the plant is pure-breeding, or homozygous, for a particular characteristic, as it then possesses two identical alleles and displays the appropriate phenotype (*Figure 5.10A*). However, Mendel showed that if two pure-breeding plants with different phenotypes are crossed then all the progeny (the F₁ generation) display the same phenotype. These F₁ plants must be heterozygous, meaning that they possess two different alleles, one for each phenotype, one allele inherited from the mother and one from the father. Mendel postulated that in this heterozygous condition one allele overrides the effects of the other allele; he therefore described the phenotype expressed in the F₁ plants as being dominant over the second, recessive phenotype (*Figure 5.10B*). This is the perfectly correct interpretation of the interaction between the pairs of alleles studied by Mendel, but we now appreciate that this simple dominant-recessive rule can be complicated by situations that he did not encounter. One of these is incomplete dominance, where the heterozygous phenotype is intermediate between the two homozygous forms. An example is when red carnations are crossed with white ones, the F₁ heterozygotes being pink. Another complication is codominance, when both alleles are detectable in the heterozygote. Codominance is the typical situation for DNA markers.

As well as discovering dominance and recessiveness, Mendel carried out additional crosses that enabled him to establish two Laws of Genetics. The First Law states that *alleles segregate randomly*. In other words, if the parent's alleles are A and a, then a member of the F₁ generation has the same chance of inheriting A as it has of inheriting a (*Figure 5.11A*). The Second Law is that *pairs of alleles segregate independently*, so that inheritance of the alleles of gene A is independent of inheritance of the alleles of gene B (*Figure 5.11B*). Because of these laws, the outcomes of genetic crosses are predictable (*Figure 5.11C*).

When Mendel's work was rediscovered in 1900, his Second Law worried the early geneticists because it was soon established that genes reside on chromosomes, and it was realized that all organisms have many more genes than chromosomes. Chromosomes are inherited as intact units, so it was reasoned that the alleles of some pairs of genes will be inherited together because they are on the same chromosome (*Figure 5.12*). This is the principle of genetic linkage, and it was quickly shown to be correct, although the results did not turn out exactly as expected. The complete linkage that had been anticipated between many pairs of genes failed to materialize. Pairs of genes were either inherited independently, as expected for genes in different chromosomes, or, if they showed linkage, then it was only partial linkage: sometimes they were inherited together and sometimes they were not (*Figure 5.13*). The resolution of this contradiction between theory and observation was the critical step in the development of genetic mapping techniques.

Partial linkage is explained by the behavior of chromosomes during meiosis

The critical breakthrough was achieved by Thomas Hunt Morgan, who made the conceptual leap between partial linkage and the behavior of chromosomes when the nucleus of a cell divides. Cytologists in the late 19th century had distinguished two types of nuclear division: mitosis and meiosis. Mitosis is more common, being the process by which the diploid nucleus of a somatic cell divides to produce two daughter nuclei, both of which are diploid (*Figure 5.14*). Approximately 10^{17} mitoses are needed to produce all the cells required during a human lifetime. Before mitosis begins, each chromosome in the nucleus is replicated, but the resulting daughter chromosomes do not immediately break away from one another. To begin with they remain attached at their centromeres and by cohesin proteins which act as 'molecular glue' holding together the arms of the replicated chromosomes (see *Figure 13.23*). The daughters do not separate until later in mitosis when the chromosomes are distributed between the two new nuclei. Obviously it is important that each of the new nuclei receives a complete set of chromosomes, and most of the intricacies of mitosis appear to be devoted to achieving this end.

Mitosis illustrates the basic events occurring during nuclear division but is not directly relevant to genetic mapping. Instead, it is the distinctive features of meiosis that interest us. Meiosis occurs only in reproductive cells, and results in a diploid cell giving rise to four haploid gametes, each of which can subsequently fuse with a gamete of the opposite sex during sexual reproduction. The fact that meiosis results in four haploid cells whereas mitosis gives rise to two

diploid cells is easy to explain: meiosis involves two nuclear divisions, one after the other, whereas mitosis is just a single nuclear division. This is an important distinction, but the critical difference between mitosis and meiosis is more subtle. Recall that in a diploid cell there are two separate copies of each chromosome (Chapter 1). We refer to these as pairs of homologous chromosomes. During mitosis, homologous chromosomes remain separate from one another, each member of the pair replicating and being passed to a daughter nucleus independently of its homolog. In meiosis, however, the pairs of homologous chromosomes are by no means independent. During meiosis I, each chromosome lines up with its homolog to form a bivalent (Figure 5.15). This occurs after each chromosome has replicated, but before the replicated structures split, so the bivalent in fact contains four chromosome copies, each of which is destined to find its way into one of the four gametes that will be produced at the end of the meiosis. Within the bivalent, the chromosome arms (the chromatids) can undergo physical breakage and exchange of segments of DNA. The process is called crossing-over or recombination and was discovered by the Belgian cytologist Janssens in 1909. This was just 2 years before Morgan started to think about partial linkage.

How did the discovery of crossing-over help Morgan explain partial linkage? To understand this we need to think about the effect that crossing-over can have on the inheritance of genes. Let us consider two genes, each of which has two alleles. We will call the first gene A and its alleles A and a, and the second gene B with alleles B and b. Imagine that the two genes are located on chromosome number 2 of *Drosophila melanogaster*, the species of fruit fly studied by Morgan. We are going to follow the meiosis of a diploid nucleus in which one copy of chromosome 2 has alleles A and B, and the second has a and b. This situation is illustrated in Figure 5.16. Consider the two alternative scenarios:

1. **A crossover does not occur between genes A and B.** If this is what happens then two of the resulting gametes will contain chromosome copies with alleles A and B, and the other two will contain a and b. In other words, two of the gametes have the genotypeAB and two have the genotype ab.

2. **A crossover does occur between genes A and B.** This leads to segments of DNA containing gene B being exchanged between homologous chromosomes. The eventual result is that each gamete has a different genotype: 1 AB, 1 aB, 1 Ab, 1 ab.

Now think about what would happen if we looked at the results of meiosis in a hundred identical cells. If crossovers never occur then the resulting gametes will have the following genotypes:

200 *AB*
200 *ab*

This is complete linkage: genes A and B behave as a single unit during meiosis. But if (as is more likely) crossovers occur between A and B in some of the nuclei, then the allele pairs will not be inherited as single units. Let us say that crossovers occur during 40 of the 100 meioses. The following gametes will result:

160 *AB*
160 *ab*
40 *Ab*
40 *aB*

The linkage is not complete, it is only partial. As well as the two **parental** genotypes (*AB*, *ab*) we see gametes with recombinant genotypes (*Ab*, *aB*). [↑ TOP](#)

From partial linkage to genetic mapping

Once Morgan had understood how partial linkage could be explained by crossing-over during meiosis he was able to devise a way of mapping the relative positions of genes on a chromosome. In fact the most important work was done not by Morgan himself, but by an undergraduate in his laboratory, Arthur Sturtevant ([Sturtevant, 1913](#)). Sturtevant assumed that crossing-over was a random event, there being an equal chance of it occurring at any position along a pair of lined-up chromatids. If this assumption is correct then two genes that are close together will be separated by crossovers less frequently than two genes that are more distant from one another. Furthermore, the frequency with which the genes are unlinked by crossovers will be directly proportional to how far apart they are on their chromosome. The recombination frequency is therefore a measure of the distance between two genes. If you work out the recombination frequencies for different pairs of genes, you can construct a map of their relative positions on the chromosome ([Figure 5.17](#)).

It turns out that Sturtevant's assumption about the randomness of crossovers was not entirely justified. Comparisons between genetic maps and the actual positions of genes on DNA molecules, as revealed by physical mapping and DNA sequencing, have shown that some regions of chromosomes, called recombination hotspots, are more likely to be involved in crossovers than others. This means that a genetic map distance does not necessarily indicate the physical distance between two markers (see [Figure 5.22](#)). Also, we now realize that a single

chromatid can participate in more than one crossover at the same time, but that there are limitations on how close together these crossovers can be, leading to more inaccuracies in the mapping procedure. Despite these qualifications, linkage analysis usually makes correct deductions about gene order, and distance estimates are sufficiently accurate to generate genetic maps that are of value as frameworks for genome sequencing projects. [↑ TOP](#)

5.2.4. Linkage analysis with different types of organism

To see how linkage analysis is actually carried out, we need to consider three quite different situations:

- Linkage analysis with species such as fruit flies and mice, with which we can carry out planned breeding experiments;
- Linkage analysis with humans, with whom we cannot carry out planned experiments but instead make use of family pedigrees;
- Linkage analysis with bacteria, which do not undergo meiosis.

Linkage analysis when planned breeding experiments are possible

The first type of linkage analysis is the modern counterpart of the method developed by Morgan and his colleagues. The method is based on analysis of the progeny of experimental crosses set up between parents of known genotypes and is, at least in theory, applicable to all eukaryotes. Ethical considerations preclude this approach in humans, and practical problems such as the length of the gestation period and the time taken for the newborn to reach maturity (and hence to participate in subsequent crosses) limit the effectiveness of the method with some animals and plants.

If we return to Figure 5.16 we see that the key to gene mapping is being able to determine the genotypes of the gametes resulting from meiosis. In a few situations this is possible by directly examining the gametes. For example, the gametes produced by some microbial eukaryotes, including the yeast *Saccharomyces cerevisiae*, can be grown into colonies of haploid cells, whose genotypes can be determined by biochemical tests. Direct genotyping of gametes is also possible with higher eukaryotes if DNA markers are used, as PCR can be carried out with the DNA from individual spermatozoa, enabling RFLPs, SSLPs and SNPs to be typed. Unfortunately, sperm typing is laborious. Routine linkage analysis with higher eukaryotes is therefore carried out not by examining the gametes directly but by determining the genotypes of

the diploid progeny that result from fusion of two gametes, one from each of a pair of parents. In other words, a genetic cross is performed.

The complication with a genetic cross is that the resulting diploid progeny are the product not of one meiosis but of two (one in each parent), and in most organisms crossover events are equally likely to occur during production of the male and female gametes. Somehow we have to be able to disentangle from the genotypes of the diploid progeny the crossover events that occurred in each of these two meioses. This means that the cross has to be set up with care. The standard procedure is to use a test cross. This is illustrated in *Figure 5.18*, Scenario 1, where we have set up a test cross to map the two genes we met earlier: gene A (alleles A and a) and gene B (alleles B and b), both on chromosome 2 of the fruit fly. The critical feature of a test cross is the genotypes of the two parents:

- One parent is a double heterozygote. This means that all four alleles are present in this parent: its genotype is AB/ab . This notation indicates that one pair of the homologous chromosomes has alleles A and B , and the other has a and b . Double heterozygotes can be obtained by crossing two pure-breeding strains, for example $AB/AB \times ab/ab$.
- The second parent is a pure-breeding double homozygote. In this parent both homologous copies of chromosome 2 are the same: in the example shown in Scenario 1 both have alleles a and b and the genotype of the parent is ab/ab .

The double heterozygote has the same genotype as the cell whose meiosis we followed in *Figure 5.16*. Our objective is therefore to infer the genotypes of the gametes produced by this parent and to calculate the fraction that are recombinants. Note that all the gametes produced by the second parent (the double homozygote) will have the genotype ab regardless of whether they are parental or recombinant gametes. Alleles a and b are both recessive, so meiosis in this parent is, in effect, invisible when the genotypes of the progeny are examined. This means that, as shown in Scenario 1 in *Figure 5.18*, the genotypes of the diploid progeny can be unambiguously converted into the genotypes of the gametes from the double heterozygous parent. The test cross therefore enables us to make a direct examination of a single meiosis and hence to calculate a recombination frequency and map distance for the two genes being studied.

Just one additional point needs to be considered. If, as in Scenario 1 in *Figure 5.18*, gene markers displaying dominance and recessiveness are used, then the double homozygous parent must have alleles for the two recessive phenotypes; however, if codominant DNA markers are

used, then the double homozygous parent can have any combination of homozygous alleles (i.e. AB/AB , Ab/Ab , aB/aB and ab/ab). Scenario 2 in [Figure 5.18](#) shows the reason for this.

Gene mapping by human pedigree analysis


With humans it is of course impossible to pre-select the genotypes of parents and set up crosses designed specifically for mapping purposes. Instead, data for the calculation of recombination frequencies have to be obtained by examining the genotypes of the members of successive generations of existing families. This means that only limited data are available, and their interpretation is often difficult because a human marriage rarely results in a convenient test cross, and often the genotypes of one or more family members are unobtainable because those individuals are dead or unwilling to cooperate.

The problems are illustrated by [Figure 5.19](#). In this example we are studying a genetic disease present in a family of two parents and six children. Genetic diseases are frequently used as gene markers in humans, the disease state being one allele and the healthy state being a second allele. The pedigree in [Figure 5.19A](#) shows us that the mother is affected by the disease, as are four of her children. We know from family accounts that the maternal grandmother also suffered from this disease, but both she and her husband - the maternal grandfather - are now dead. We can include them in the pedigree, with slashes indicating that they are dead, but we cannot obtain any further information on their genotypes. Our aim is to map the position of the gene for the genetic disease. For this purpose we are studying its linkage to a microsatellite marker M, four alleles of which - M_1 , M_2 , M_3 and M_4 - are present in the living family members. The question is, how many of the children are recombinants?

If we look at the genotypes of the six children we see that numbers 1, 3 and 4 have the disease allele and the microsatellite allele M_1 . Numbers 2 and 5 have the healthy allele and M_2 . We can therefore construct two alternative hypotheses. The first is that the two copies of the relevant homologous chromosomes in the mother have the genotypes *Disease- M_1* and *Healthy- M_2* ; therefore children 1, 2, 3, 4 and 5 have parental genotypes and child 6 is the one and only recombinant ([Figure 5.19B](#)). This would suggest that the disease gene and the microsatellite are relatively closely linked and that crossovers between them occur infrequently. The alternative hypothesis is that the mother's chromosomes have the genotypes *Healthy- M_1* and *Disease- M_2* ; this would mean that children 1–5 are recombinants, and child 6 has the parental genotype. This

would mean that the gene and microsatellite are relatively far apart on the chromosome. We cannot determine which of these hypotheses is correct: the data are frustratingly ambiguous.

The most satisfying solution to the problem posed by the pedigree in [Figure 5.19](#) would be to know the genotype of the grandmother. Let us pretend that this is a soap opera family and that the grandmother is not really dead. To everyone's surprise she reappears just in time to save the declining audience ratings. Her genotype for microsatellite M turns out to be M_1M_5 ([Figure 5.19C](#)). This tells us that the disease allele is on the same chromosome as M_1 . We can therefore conclude with certainty that Hypothesis 1 is correct and that only child 6 is a recombinant.

Resurrection of key individuals is not usually an option open to real-life geneticists, although DNA can be obtained from old pathology specimens such as slides and Guthrie cards. Imperfect pedigrees are analyzed statistically, using a measure called the lod score ([Morton, 1955](#)). This stands for logarithm of the odds that the genes are linked and is used primarily to determine if the two markers being studied lie on the same chromosome, in other words if the genes are linked or not. If the lod analysis establishes linkage then it can also provide a measure of the most likely recombination frequency. Ideally the available data will derive from more than one pedigree, increasing the confidence in the result. The analysis is less ambiguous for families with larger numbers of children, and, as we saw in [Figure 5.19](#), it is important that the members of at least three generations can be genotyped. For this reason, family collections have been established, such as the one maintained by the Centre d'Études du Polymorphisme Humaine (CEPH) in Paris ( [Dausset et al., 1990](#)). The CEPH collection contains cultured cell lines from families in which all four grandparents as well as at least eight second-generation children could be sampled. This collection is available for DNA marker mapping by any researcher who agrees to submit the resulting data to the central CEPH database. [↑ TOP](#)

Genetic mapping in bacteria

The final type of genetic mapping that we must consider is the strategy used with bacteria. The main difficulty that geneticists faced when trying to develop genetic mapping techniques for bacteria is that these organisms are normally haploid, and so do not undergo meiosis. Some other way therefore had to be devised to induce crossovers between homologous segments of bacterial DNA. The answer was to make use of three natural methods that exist for transferring pieces of DNA from one bacterium to another ([Figure 5.20](#)):


- In conjugation two bacteria come into physical contact and one bacterium (the donor) transfers DNA to the second bacterium (the recipient). The transferred DNA can be a copy of some or possibly all of the donor cell's chromosome, or it could be a segment of chromosomal DNA - up to 1 Mb in length - integrated in a plasmid (Section 2.1.2). The latter is called episome transfer.
- Transduction involves transfer of a small segment of DNA - up to 50 kb or so - from donor to recipient via a bacteriophage.
- In transformation the recipient cell takes up from its environment a fragment of DNA, rarely longer than 50 kb, released from a donor cell.

After transfer, a double crossover must occur so that the DNA from the donor bacterium is integrated into the recipient cell's chromosome (Figure 5.21A). If this does not occur then the transferred DNA is lost when the recipient cell divides. The only exception is after episome transfer, plasmids being able to propagate independently of the host chromosome.

Biochemical markers are invariably used, the dominant or **wild-type** phenotype being possession of a biochemical characteristic (e.g. ability to synthesize tryptophan) and the recessive phenotype being the complementary characteristic (e.g. inability to synthesize tryptophan). The gene transfer is usually set up between a donor strain that possesses the wild-type alleles and a recipient with the recessive alleles, transfer into the recipient strain being monitored by looking for acquisition of the biochemical function(s) specified by the genes being studied. The precise details of the mapping procedure depend on the type of gene transfer that is being used. In conjugation mapping the donor DNA is transferred as a continuous thread into the recipient, and gene positions are mapped by timing the entry of the wild-type alleles into the recipient (Figure 5.21B). Transduction and transformation mapping enable genes that are relatively close together to be mapped, because the transferred DNA segment is short (< 50 kb), so the probability of two genes being transferred together depends on how close together they are on the bacterial chromosome (Figure 5.21C).

5.3. Physical Mapping

A map generated by genetic techniques is rarely sufficient for directing the sequencing phase of a genome project. This is for two reasons:

- ***The resolution of a genetic map depends on the number of crossovers that have been scored.*** This is not a major problem for microorganisms because these can be obtained in huge numbers, enabling many crossovers to be studied, resulting in a highly detailed genetic map in which the markers are just a few kb apart. For example, when the *Escherichia coli* genome sequencing project began in 1990, the latest genetic map for this organism comprised over 1400 markers, an average of one per 3.3 kb. This was sufficiently detailed to direct the sequencing program without the need for extensive physical mapping. Similarly, the *Saccharomyces cerevisiae* project was supported by a fine-scale genetic map (approximately 1150 genetic markers, on average one per 10 kb). The problem with humans and most other eukaryotes is that it is simply not possible to obtain large numbers of progeny, so relatively few meioses can be studied and the resolving power of linkage analysis is restricted. This means that genes that are several tens of kb apart may appear at the same position on the genetic map.
- ***Genetic maps have limited accuracy.*** We touched on this point in [Section 5.2.3](#) when we assessed Sturtevant's assumption that crossovers occur at random along chromosomes. This assumption is only partly correct because the presence of recombination hotspots means that crossovers are more likely to occur at some points rather than at others. The effect that this can have on the accuracy of a genetic map was illustrated in 1992 when the complete sequence for *S. cerevisiae* chromosome III was published ( [Oliver et al., 1992](#)), enabling the first direct comparison to be made between a genetic map and the actual positions of markers as shown by DNA sequencing ([Figure 5.22](#)). There were considerable discrepancies, even to the extent that one pair of genes had been ordered incorrectly by genetic analysis. Bear in mind that *S. cerevisiae* is one of the two eukaryotes (fruit fly is the second) whose genomes have been subjected to intensive genetic mapping. If the yeast genetic map is inaccurate then how precise are the genetic maps of organisms subjected to less detailed analysis?

These two limitations of genetic mapping mean that for most eukaryotes a genetic map must be checked and supplemented by alternative mapping procedures before large-scale DNA sequencing begins. A plethora of physical mapping techniques has been developed to address this problem, the most important being:

- **Restriction mapping**, which locates the relative positions on a DNA molecule of the recognition sequences for restriction endonucleases;
- **Fluorescent *in situ* hybridization (FISH)**, in which marker locations are mapped by hybridizing a probe containing the marker to intact chromosomes;
- **Sequence tagged site (STS) mapping**, in which the positions of short sequences are mapped by PCR and/or hybridization analysis of genome fragments.

5.3.1. Restriction mapping

Genetic mapping using RFLPs as DNA markers can locate the positions of polymorphic restriction sites within a genome ([Section 5.2.2](#)), but very few of the restriction sites in a genome are polymorphic, so many sites are not mapped by this technique ([Figure 5.23](#)). Could we increase the marker density on a genome map by using an alternative method to locate the positions of some of the non-polymorphic restriction sites? This is what restriction mapping achieves, although in practice the technique has limitations which mean that it is applicable only to relatively small DNA molecules. We will look first at the technique and then consider its relevance to genome mapping.

The basic methodology for restriction mapping

The simplest way to construct a restriction map is to compare the fragment sizes produced when a DNA molecule is digested with two different restriction enzymes that recognize different target sequences. An example using the restriction enzymes *EcoRI* and *BamHI* is shown in [Figure 5.24](#) . First, the DNA molecule is digested with just one of the enzymes and the sizes of the resulting fragments are measured by agarose gel electrophoresis. Next, the molecule is digested with the second enzyme and the resulting fragments again sized in an agarose gel. The results so far enable the number of restriction sites for each enzyme to be worked out, but do not allow their relative positions to be determined. Additional information is therefore obtained by cutting the DNA molecule with both enzymes together. In the example shown in [Figure 5.24](#) , this double

restriction enables three of the sites to be mapped. However, a problem arises with the larger *EcoRI* fragment because this contains two *Bam*HI sites and there are two alternative possibilities for the map location of the outer one of these. The problem is solved by going back to the original DNA molecule and treating it again with *Bam*HI on its own, but this time preventing the digestion from going to completion by, for example, incubating the reaction for only a short time or using a suboptimal incubation temperature. This is called a partial restriction and leads to a more complex set of products, the complete restriction products now being supplemented with partially restricted fragments that still contain one or more uncut *Bam*HI sites. In the example shown in *Figure 5.24*, the size of one of the partial restriction fragments is diagnostic and the correct map can be identified.

A partial restriction usually gives the information needed to complete a map, but if there are many restriction sites then this type of analysis becomes unwieldy, simply because there are so many different fragments to consider. An alternative strategy is simpler because it enables the majority of the fragments to be ignored. This is achieved by attaching a radioactive or other type of marker to each end of the starting DNA molecule before carrying out the partial digestion. The result is that many of the partial restriction products become 'invisible' because they do not contain an end-fragment and so do not show up when the agarose gel is screened for labeled products. The sizes of the partial restriction products that are visible enable unmapped sites to be positioned relative to the ends of the starting molecule. [↑ TOP](#)

The scale of restriction mapping is limited by the sizes of the restriction fragments

Restriction maps are easy to generate if there are relatively few cut sites for the enzymes being used. However, as the number of cut sites increases, so also do the numbers of single, double and partial restriction products whose sizes must be determined and compared in order for the map to be constructed. Computer analysis can be brought into play but problems still eventually arise. A stage will be reached when a digest contains so many fragments that individual bands merge on the agarose gel, increasing the chances of one or more fragments being measured incorrectly or missed out entirely. If several fragments have similar sizes then even if they can all be identified, it may not be possible to assemble them into an unambiguous map.

Restriction mapping is therefore more applicable to small rather than large molecules, with the upper limit for the technique depending on the frequency of the restriction sites in the molecule being mapped. In practice, if a DNA molecule is less than 50 kb in length it is usually possible to

construct an unambiguous restriction map for a selection of enzymes with six-nucleotide recognition sequences. Fifty kb is of course way below the minimum size for bacterial or eukaryotic chromosomes, although it does cover a few viral and organelle genomes, and whole-genome restriction maps have indeed been important in directing sequencing projects with these small molecules. Restriction maps are equally useful after bacterial or eukaryotic genomic DNA has been cloned, if the cloned fragments are less than 50 kb, because a detailed restriction map can then be built up as a preliminary to sequencing the cloned region. This is an important application of restriction mapping in sequencing projects with large genomes, but is there any possibility of using restriction analysis for the more general mapping of entire genomes larger than 50 kb?

The answer is a qualified 'yes', because the limitations of restriction mapping can be eased slightly by choosing enzymes expected to have infrequent cut sites in the target DNA molecule. These 'rare cutters' fall into two categories:



- ***Enzymes with seven- or eight-nucleotide recognition sequences.*** A few restriction enzymes cut at seven- or eight-nucleotide recognition sequences. Examples are *SapI* (5'-GCTCTTC-3') and *SgfI* (5'-GCGATCGC-3'). The seven-nucleotide enzymes would be expected, on average, to cut a DNA molecule with a GC content of 50% once every $4^7 = 16\,384$ bp. The eight-nucleotide enzymes should cut once every $4^8 = 65\,536$ bp. These figures compare with $4^6 = 4096$ bp for six-nucleotide enzymes such as *BamHI* and *EcoRI*. Seven- and eight-nucleotide cutters are often used in restriction mapping of large molecules but the approach is not as useful as it might be simply because not many of these enzymes are known.
- ***Enzymes whose recognition sequences contain motifs that are rare in the target DNA.*** Genomic DNA molecules do not have random sequences and some are significantly deficient in certain motifs. For example, the sequence 5'-CG-3' is rare in human DNA because human cells possess an enzyme that adds a methyl group to carbon 5 of the C nucleotide in this sequence. The resulting 5-methylcytosine is unstable and tends to undergo deamination to give thymine ([Figure 5.25](#)). The consequence is that during human evolution many of the 5'-CG-3' sequences that were originally in our genome have become converted to 5'-TG-3'. Restriction enzymes that recognize a site containing 5'-CG-3' therefore cut human DNA relatively infrequently. Examples are *SmaI* (5'-

CCCGGG-3'), which cuts human DNA on average once every 78 kb, and *Bss*HII (5'-GCGCGC-3') which cuts once every 390 kb. Note that *Not*I, an eight-nucleotide cutter, also targets 5'-CG-3' sequences (recognition sequence 5'-GCGGCCGC-3') and cuts human DNA very rarely - approximately once every 10 Mb.


The potential of restriction mapping is therefore increased by using rare cutters. It is still not possible to construct restriction maps of the genomes of animals and plants, but it is feasible to use the technique with large cloned fragments, and the smaller DNA molecules of prokaryotes and lower eukaryotes such as yeast and fungi.



If a rare cutter is used then it may be necessary to employ a special type of agarose gel electrophoresis to study the resulting restriction fragments. This is because the relationship between the length of a DNA molecule and its migration rate in an electrophoresis gel is not linear, the resolution decreasing as the molecules get longer (*Figure 5.26A*). This means that it is not possible to separate molecules more than about 50 kb in length because all of these longer molecules run as a single slowly migrating band in a standard agarose gel. To separate them it is necessary to replace the linear electric field used in conventional gel electrophoresis with a more complex field. An example is provided by orthogonal field alternation gel electrophoresis (OFAGE), in which the electric field alternates between two pairs of electrodes, each positioned at an angle of 45° to the length of the gel (*Figure 5.26B*). The DNA molecules still move down through the gel, but each change in the field forces the molecules to realign. Shorter molecules realign more quickly than longer ones and so migrate more rapidly through the gel. The overall result is that molecules much longer than those separated by conventional gel electrophoresis can be resolved.

Direct examination of DNA molecules for restriction sites


It is also possible to use methods other than electrophoresis to map restriction sites in DNA molecules. With the technique called optical mapping ( Schwartz *et al.*, 1993), restriction sites are directly located by looking at the cut DNA molecules with a microscope (*Figure 5.27*). The DNA must first be attached to a glass slide in such a way that the individual molecules become stretched out, rather than clumped together in a mass. There are two ways of doing this: gel stretching and molecular combing. To prepare gel-stretched DNA fibers ( Schwartz *et al.*, 1993), chromosomal DNA is suspended in molten agarose and placed on a microscope slide. As the gel cools and solidifies, the DNA molecules become extended (*Figure 5.28A*). To utilize gel

stretching in optical mapping, the microscope slide onto which the molten agarose is placed is first coated with a restriction enzyme. The enzyme is inactive at this stage because there are no magnesium ions, which the enzyme needs in order to function. Once the gel has solidified it is washed with a solution containing magnesium chloride, which activates the restriction enzyme. A fluorescent dye is added, such as DAPI (4,6-diamino-2-phenylindole dihydrochloride), which stains the DNA so that the fibers can be seen when the slide is examined with a high-power fluorescence microscope. The restriction sites in the extended molecules gradually become gaps as the degree of fiber extension is reduced by the natural springiness of the DNA, enabling the relative positions of the cuts to be recorded.

In molecular combing (Michalet *et al.*, 1997), the DNA fibers are prepared by dipping a silicone-coated cover slip into a solution of DNA, leaving it for 5 minutes (during which time the DNA molecules attach to the cover slip by their ends), and then removing the slip at a constant speed of 0.3 mm s^{-1} (*Figure 5.28B*). The force required to pull the DNA molecules through the meniscus causes them to line up. Once in the air, the surface of the cover slip dries, retaining the DNA molecules as an array of parallel fibers.

Optical mapping was first applied to large DNA fragments cloned in YAC and BAC vectors (*Section 4.2.1*). More recently, the feasibility of using this technique with genomic DNA has been proven with studies of a 1-Mb chromosome of the malaria parasite *Plasmodium falciparum* (Jing *et al.*, 1999), and the two chromosomes and single megaplasmid of the bacterium *Deinococcus radiodurans* (Lin *et al.*, 1999).

5.3.2. Fluorescent *in situ* hybridization (FISH)

The optical mapping method described above provides a link to the second type of physical mapping procedure that we will consider - FISH (Heiskanen *et al.*, 1996). As in optical mapping, FISH enables the position of a marker on a chromosome or extended DNA molecule to be directly visualized. In optical mapping the marker is a restriction site and it is visualized as a gap in an extended DNA fiber. In FISH, the marker is a DNA sequence that is visualized by hybridization with a fluorescent probe.

In situ hybridization with radioactive or fluorescent probes

In situ hybridization is a version of hybridization analysis (*Section 4.1.2*) in which an intact chromosome is examined by probing it with a labeled DNA molecule. The position on the

chromosome at which hybridization occurs provides information about the map location of the DNA sequence used as the probe (*Figure 5.29*). For the method to work, the DNA in the chromosome must be made single stranded ('denatured') by breaking the base pairs that hold the double helix together. Only then will the chromosomal DNA be able to hybridize with the probe. The standard method for denaturing chromosomal DNA without destroying the morphology of the chromosome is to dry the preparation onto a glass microscope slide and then treat with formamide.

In the early versions of *in situ* hybridization the probe was radioactively labeled but this procedure was unsatisfactory because it is difficult to achieve both sensitivity and resolution with a radioactive label, two critical requirements for successful *in situ* hybridization. Sensitivity requires that the radioactive label has a high emission energy (an example of such a radiolabel is ^{32}P), but if the radiolabel has a high emission energy then it scatters its signal and so gives poor resolution. High resolution is possible if a radiolabel with low emission energy, such as ^3H , is used, but these have such low sensitivity that lengthy exposures are needed, leading to a high background and difficulties in discerning the genuine signal.

These problems were solved in the late 1980s by the development of non-radioactive fluorescent DNA labels. These labels combine high sensitivity with high resolution and are ideal for *in situ* hybridization. Fluorolabels with different colored emissions have been designed, making it possible to hybridize a number of different probes to a single chromosome and distinguish their individual hybridization signals, thus enabling the relative positions of the probe sequences to be mapped. To maximize sensitivity, the probes must be labeled as heavily as possible, which in the past has meant that they must be quite lengthy DNA molecules - usually cloned DNA fragments of at least 40 kb. This requirement is less important now that techniques for achieving heavy labeling with shorter molecules have been developed. As far as the construction of a physical map is concerned, a cloned DNA fragment can be looked upon as simply another type of marker, although in practice the use of clones as markers adds a second dimension because the cloned DNA is the material from which the DNA sequence is determined. Mapping the positions of clones therefore provides a direct link between a genome map and its DNA sequence.

If the probe is a long fragment of DNA then one potential problem, at least with higher eukaryotes, is that it is likely to contain examples of repetitive DNA sequences (*Section 2.4*) and so may hybridize to many chromosomal positions, not just the specific point to which it is

perfectly matched. To reduce this non-specific hybridization, the probe, before use, is mixed with unlabeled DNA from the organism being studied. This DNA can simply be total nuclear DNA (i.e. representing the entire genome) but it is better if a fraction enriched for repeat sequences is used. The idea is that the unlabeled DNA hybridizes to the repetitive DNA sequences in the probe, blocking these so that the subsequent *in situ* hybridization is driven wholly by the unique sequences (Lichter *et al.*, 1990). Non-specific hybridization is therefore reduced or eliminated entirely (*Figure 5.30*).

FISH in action

FISH was originally used with metaphase chromosomes (Section 2.2.1). These chromosomes, prepared from nuclei that are undergoing division, are highly condensed and each chromosome in a set takes up a recognizable appearance, characterized by the position of its centromere and the banding pattern that emerges after the chromosome preparation is stained (see *Figure 2.8*). With metaphase chromosomes, a fluorescent signal obtained by FISH is mapped by measuring its position relative to the end of the short arm of the chromosome (the *FLpter value*). A disadvantage is that the highly condensed nature of metaphase chromosomes means that only low-resolution mapping is possible, two markers having to be at least 1 Mb apart to be resolved as separate hybridization signals (Trask *et al.*, 1991). This degree of resolution is insufficient for the construction of useful chromosome maps, and the main application of metaphase FISH has been in determining the chromosome on which a new marker is located, and providing a rough idea of its map position, as a preliminary to finer scale mapping by other methods.

For several years these 'other methods' did not involve any form of FISH, but since 1995 a range of higher resolution FISH techniques has been developed. With these techniques, higher resolution is achieved by changing the nature of the chromosomal preparation being studied. If metaphase chromosomes are too condensed for fine-scale mapping then we must use chromosomes that are more extended. There are two ways of doing this (Heiskanen *et al.*, 1996):

- **Mechanically stretched chromosomes** can be obtained by modifying the preparative method used to isolate chromosomes from metaphase nuclei. The inclusion of a centrifugation step generates shear forces which can result in the chromosomes becoming stretched to up to 20 times their normal length. Individual chromosomes are still recognizable and FISH signals can be mapped in the same way as with normal metaphase

chromosomes. The resolution is significantly improved and markers that are 200–300 kb apart can be distinguished.

- **Non-metaphase chromosomes** can be used because it is only during metaphase that chromosomes are highly condensed: at other stages of the cell cycle the chromosomes are naturally unpacked. Attempts have been made to use prophase nuclei (see *Figure 5.14*) because in these the chromosomes are still sufficiently condensed for individual ones to be identified. In practice, however, these preparations provide no advantage over mechanically stretched chromosomes. Interphase chromosomes are more useful because this stage of the cell cycle (between nuclear divisions) is when the chromosomes are most unpacked. Resolution down to 25 kb is possible, but chromosome morphology is lost so there are no external reference points against which to map the position of the probe. This technique is therefore used after preliminary map information has been obtained, usually as a means of determining the order of a series of markers in a small region of a chromosome.

Interphase chromosomes contain the most unpacked of all cellular DNA molecules. To improve the resolution of FISH to better than 25 kb it is therefore necessary to abandon intact chromosomes and instead use purified DNA. This approach, called fiber-FISH, makes use of DNA prepared by gel stretching or molecular combing (see *Figure 5.28*) and can distinguish markers that are less than 10 kb apart.

5.3.3. Sequence tagged site (STS) mapping

To generate a detailed physical map of a large genome we need, ideally, a high-resolution mapping procedure that is rapid and not technically demanding. Neither of the two techniques that we have considered so far - restriction mapping and FISH - meets these requirements. Restriction mapping is rapid, easy, and provides detailed information, but it cannot be applied to large genomes. FISH can be applied to large genomes, and modified versions such as fiber-FISH can give high-resolution data, but FISH is difficult to carry out and data accumulation is slow, map positions for no more than three or four markers being obtained in a single experiment. If detailed physical maps are to become a reality then we need a more powerful technique.

At present the most powerful physical mapping technique, and the one that has been responsible for generation of the most detailed maps of large genomes, is STS mapping. A sequence tagged site or **STS** is simply a short DNA sequence, generally between 100 and 500 bp in length, that is


easily recognizable and occurs only once in the chromosome or genome being studied. To map a set of STSs, a collection of overlapping DNA fragments from a single chromosome or from the entire genome is needed. In the example shown in [Figure 5.31](#) a fragment collection has been prepared from a single chromosome, with each point along the chromosome represented on average five times in the collection. The data from which the map will be derived are obtained by determining which fragments contain which STSs. This can be done by hybridization analysis but PCR is generally used because it is quicker and has proven to be more amenable to automation. The chances of two STSs being present on the same fragment will, of course, depend on how close together they are in the genome. If they are very close then there is a good chance that they will always be on the same fragment; if they are further apart then sometimes they will be on the same fragment and sometimes they will not ([Figure 5.31](#)). The data can therefore be used to calculate the distance between two markers, in a manner analogous to the way in which map distances are determined by linkage analysis ([Section 5.2.3](#)). Remember that in linkage analysis a map distance is calculated from the frequency at which crossovers occur between two markers. STS mapping is essentially the same, except that each map distance is based on the frequency at which *breaks* occur between two markers.

The description of STS mapping given above leaves out some critical questions: What exactly is an STS? How is the DNA fragment collection obtained?

Any unique DNA sequence can be used as an STS

To qualify as an STS, a DNA sequence must satisfy two criteria. The first is that its sequence must be known, so that a PCR assay can be set up to test for the presence or absence of the STS on different DNA fragments. The second requirement is that the STS must have a unique location in the chromosome being studied, or in the genome as a whole if the DNA fragment set covers the entire genome. If the STS sequence occurs at more than one position then the mapping data will be ambiguous. Care must therefore be taken to ensure that STSs do not include sequences found in repetitive DNA.

These are easy criteria to satisfy and STSs can be obtained in many ways, the most common sources being **expressed sequence tags (ESTs)**, **SSLPs**, and **random genomic sequences**.


- Expressed sequence tags (ESTs). These are short sequences obtained by analysis of cDNA clones ( [Marra et al., 1998](#)). Complementary DNA is prepared by converting an mRNA preparation into double-stranded DNA ([Figure 5.32](#)). Because the mRNA in a

cell is derived from protein-coding genes, cDNAs and the ESTs obtained from them represent the genes that were being expressed in the cell from which the mRNA was prepared. ESTs are looked upon as a rapid means of gaining access to the sequences of important genes, and they are valuable even if their sequences are incomplete. An EST can also be used as an STS, assuming that it comes from a unique gene and not from a member of a gene family in which all the genes have the same or very similar sequences.

- **SSLPs.** In [Section 5.2.2](#) we examined the use of microsatellites and other SSLPs in genetic mapping. SSLPs can also be used as STSs in physical mapping. SSLPs that are polymorphic and have already been mapped by linkage analysis are particularly valuable as they provide a direct connection between the genetic and physical maps.
- **Random genomic sequences.** These are obtained by sequencing random pieces of cloned genomic DNA, or simply by downloading sequences that have been deposited in the databases.

Fragments of DNA for STS mapping


The second component of an STS mapping procedure is the collection of DNA fragments spanning the chromosome or genome being studied. This collection is sometimes called the mapping reagent and at present there are two ways in which it can be assembled: as a clone library and as a panel of radiation hybrids. We will consider radiation hybrids first.

A radiation hybrid is a rodent cell that contains fragments of chromosomes from a second organism ( McCarthy, 1996). The technology was first developed in the 1970s when it was discovered that exposure of human cells to X-ray doses of 3000–8000 rads causes the chromosomes to break up randomly into fragments, larger X-ray doses producing smaller fragments ([Figure 5.33A](#)). This treatment is of course lethal for the human cells, but the chromosome fragments can be propagated if the irradiated cells are subsequently fused with non-irradiated hamster or other rodent cells. Fusion is stimulated either chemically with polyethylene glycol or by exposure to Sendai virus ([Figure 5.33B](#)). Not all of the hamster cells take up chromosome fragments so a means of identifying the hybrids is needed. The routine selection process is to use a hamster cell line that is unable to make either thymidine kinase (TK) or hypoxanthine phosphoribosyl transferase (HPRT), deficiencies in either of these two enzymes being lethal when the cells are grown in a medium containing a mixture of hypoxanthine,

aminopterin and thymidine (HAT medium). After fusion, the cells are placed in HAT medium. Those that grow are hybrid hamster cells that have acquired human DNA fragments that include genes for the human TK and HPRT enzymes, which are synthesized inside the hybrids, enabling these cells to grow in the selective medium. The treatment results in hybrid cells that contain a random selection of human DNA fragments inserted into the hamster chromosomes. Typically the fragments are 5–10 Mb in size, with each cell containing fragments equivalent to 15–35% of the human genome. The collection of cells is called a radiation hybrid panel and can be used as a mapping reagent in STS mapping, provided that the PCR assay used to identify the STS does not amplify the equivalent region of DNA from the hamster genome.

A second type of radiation hybrid panel, containing DNA from just one human chromosome, can be constructed if the cell line that is irradiated is not a human one but a second type of rodent hybrid. Cytogeneticists have developed a number of rodent cell lines in which a single human chromosome is stably propagated in the rodent nucleus. If a cell line of this type is irradiated and fused with hamster cells, then the hybrid hamster cells obtained after selection will contain either human or mouse chromosome fragments, or a mixture of both. The ones containing human DNA can be identified by probing with a human-specific genome-wide repeat sequence, such as the short interspersed nuclear element (SINE) called Alu ([Section 2.4.2](#)), which has a copy number of just over 1 million (see [Table 1.2](#)) and so occurs on average once every 4 kb in the human genome. Only cells containing human DNA will hybridize to Alu probes, enabling the uninteresting mouse hybrids to be discarded and STS mapping to be directed at the cells containing human chromosome fragments.

Radiation hybrid mapping of the human genome was initially carried out with chromosome-specific rather than whole-genome panels because it was thought that fewer hybrids would be needed to map a single chromosome than would be needed to map the entire genome. It turns out that a high-resolution map of a single human chromosome requires a panel of 100–200 hybrids, which is about the most that can be handled conveniently in a PCR screening program. But whole-genome and single-chromosome panels are constructed differently, the former involving irradiation of just human DNA, and the latter requiring irradiation of a mouse cell containing much mouse DNA and relatively little human DNA. This means that the human DNA content per hybrid is much lower in a single-chromosome panel than in a whole-genome panel. It transpires that detailed mapping of the entire human genome is possible with fewer than 100

whole-genome radiation hybrids, so whole-genome mapping is no more difficult than single-chromosome mapping. Once this was realized, whole-genome radiation hybrids became a central component of the mapping phase of the Human Genome Project ([Section 6.3.1](#)). Whole-genome libraries are also being used for STS mapping of other mammalian genomes and for those of the zebra fish and the chicken ( [McCarthy, 1996](#)). [↑ TOP](#)

A clone library can also be used as the mapping reagent for STS analysis

A preliminary to the sequencing phase of a genome project is to break the genome or isolated chromosomes into fragments and to clone each one in a high-capacity vector, one able to handle large fragments of DNA ([Section 4.2.1](#)). This results in a clone library, a collection of DNA fragments, which, in this case, have an average size of several hundred kb. As well as supporting the sequencing work, this type of clone library can also be used as a mapping reagent in STS analysis.

As with radiation hybrid panels, a clone library can be prepared from genomic DNA, and so represents the entire genome, or a chromosome-specific library can be made if the starting DNA comes from just one type of chromosome. The latter is possible because individual chromosomes can be separated by [flow cytometry](#). To carry out this technique, dividing cells (ones with condensed chromosomes) are carefully broken open so that a mixture of intact chromosomes is obtained. The chromosomes are then stained with a fluorescent dye. The amount of dye that a chromosome binds depends on its size, so larger chromosomes bind more dye and fluoresce more brightly than smaller ones. The chromosome preparation is diluted and passed through a fine aperture, producing a stream of droplets, each one containing a single chromosome. The droplets pass through a detector that measures the amount of fluorescence, and hence identifies which droplets contain the particular chromosome being sought. An electric charge is applied to these drops, and no others ([Figure 5.34](#)), enabling the droplets containing the desired chromosome to be deflected and separated from the rest. What if two different chromosomes have similar sizes, as is the case with human chromosomes 21 and 22? These can usually be separated if the dye that is used is not one that binds non-specifically to DNA, but instead has a preference for AT- or GC-rich regions. Examples of such dyes are Hoechst 33258 and chromomycin A₃, respectively. Two chromosomes that are the same size rarely have identical GC contents, and so can be distinguished by the amounts of AT- or GC-specific dye that they bind.

Compared with radiation hybrid panels, clone libraries have one important advantage for STS mapping. This is the fact that the individual clones can subsequently provide the DNA that is actually sequenced. The data resulting from STS analysis, from which the physical map is generated, can equally well be used to determine which clones contain overlapping DNA fragments, enabling a clone contig to be built up (*Figure 5.35* ; for other methods for assembling clone contigs see *Section 6.2.2*). This assembly of overlapping clones can be used as the base material for a lengthy, continuous DNA sequence, and the STS data can later be used to anchor this sequence precisely onto the physical map. If the STSs also include SSLPs that have been mapped by genetic linkage analysis then the DNA sequence, physical map and genetic map can all be integrated

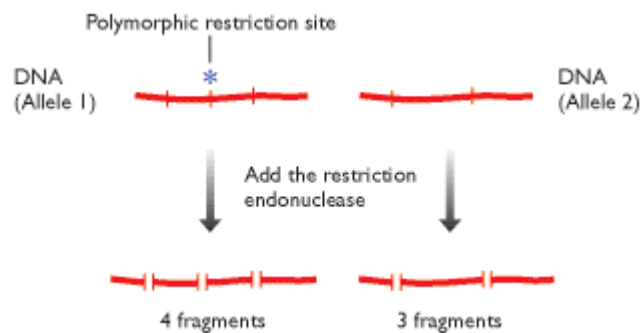


Figure 5.4. A restriction fragment length polymorphism (RFLP). The DNA molecule on the left has a polymorphic restriction site (marked with the asterisk) that is not present in the molecule on the right. The RFLP is revealed after treatment with the restriction enzyme because one of the molecules is cut into four fragments whereas the other is cut into three fragments.

Genome mapping

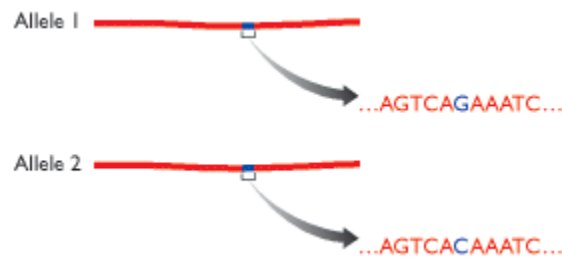


Figure 5.7. A single nucleotide polymorphism (SNP).

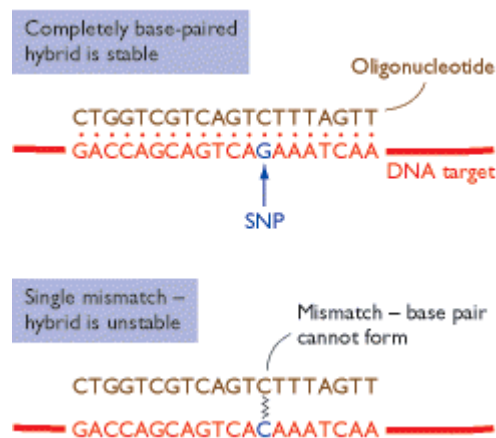


Figure 5.8. Oligonucleotide hybridization is very specific. Under highly stringent hybridization conditions, a stable hybrid occurs only if the oligonucleotide is able to form a completely base-paired structure with the target DNA. If there is a single mismatch then the hybrid does not form. To achieve this level of stringency, the incubation temperature must be just below the melting temperature or T_m of the oligonucleotide. At temperatures above the T_m , even the fully base-paired hybrid is unstable. At more than 5 °C below the T_m , mismatched hybrids might be stable. The T_m for the oligonucleotide shown in the figure would be about 58 °C. The T_m in °C is calculated from the formula $T_m = (4 \times \text{number of G and C nucleotides}) + (2 \times \text{number of A and T nucleotides})$. This formula gives a rough indication of the T_m for oligonucleotides of 15–30 nucleotides in length.

Genome mapping

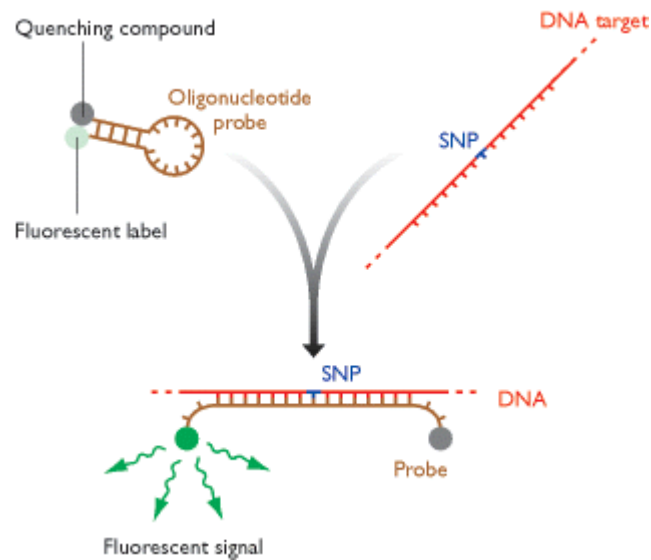


Figure 5.9. One way of detecting an SNP by solution hybridization. The oligonucleotide probe has two end-labels. One of these is a fluorescent dye and the other is a quenching compound. The two ends of the oligonucleotide base-pair to one another, so the fluorescent signal is quenched. When the probe hybridizes to its target DNA, the ends of the molecule become separated, enabling the fluorescent dye to emit its signal. The two labels are called 'molecular beacons'.

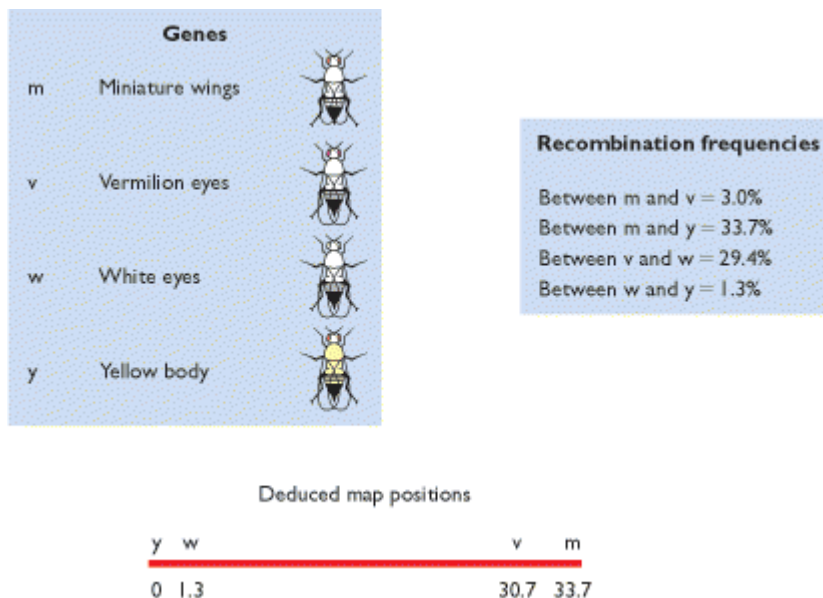
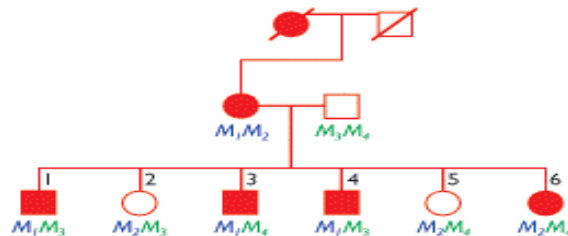


Figure 5.17. Working out a genetic map from recombination frequencies. The example is taken from the original experiments carried out with fruit flies by Arthur Sturtevant. All four genes are on the X chromosome of the fruit fly. Recombination frequencies between the genes are shown, along with their deduced map positions.

(A) The pedigree



(B) Possible interpretations of the pedigree

MOTHER'S CHROMOSOMES		
	Hypothesis 1	Hypothesis 2
	<u>Disease M_1</u>	<u>Healthy M_1</u>
	<u>Healthy M_2</u>	<u>Disease M_2</u>
CHILD 1	<u>Disease M_1</u>	Recombinant
CHILD 2	<u>Healthy M_2</u>	Recombinant
CHILD 3	<u>Disease M_1</u>	Recombinant
CHILD 4	<u>Disease M_1</u>	Recombinant
CHILD 5	<u>Healthy M_2</u>	Recombinant
CHILD 6	<u>Disease M_2</u>	Parental
Recombination frequency	1/6 = 16.7%	5/6 = 83.3%

(C) Resurrection of the maternal grandmother

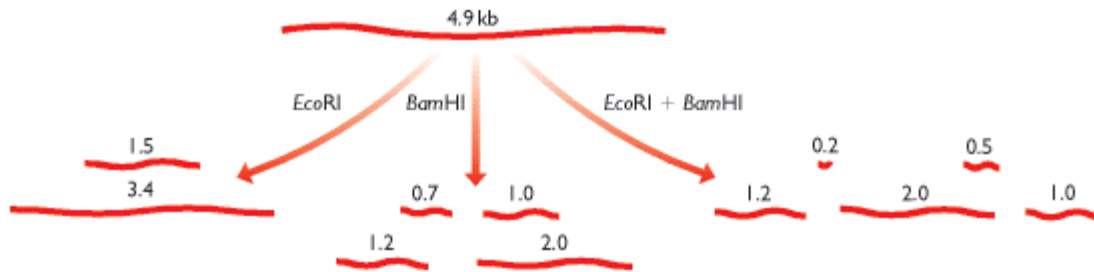


KEY				
○ Unaffected female	● Affected female	□ Unaffected male	■ Affected male	/ Dead

Figure 5.19. An example of human pedigree analysis. (A) The pedigree shows inheritance of a genetic disease in a family of two living parents and six children, with information about the maternal grandparents available from family records. The disease allele (closed symbols) is dominant over the healthy allele (open symbols). The objective is to determine the degree of

linkage between the disease gene and the microsatellite M by typing the alleles for this microsatellite (M_1 , M_2 , etc.) in living members of the family. (B) The pedigree can be interpreted in two different ways: Hypothesis 1 gives a low recombination frequency and indicates that the disease gene is tightly linked to microsatellite M; Hypothesis 2 suggests that the gene and microsatellite are much less closely linked. In (C), the issue is resolved by the reappearance of the maternal grandmother, whose microsatellite genotype is consistent only with Hypothesis 1. See the text for more details.

Genome mapping



INTERPRETATION OF THE DOUBLE RESTRICTION

Fragments

Conclusions

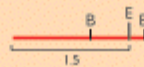
0.2 kb, 0.5 kb

These must derive from the 0.7 kb BamHI fragment, which therefore has an internal EcoRI site:



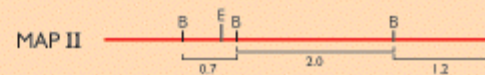
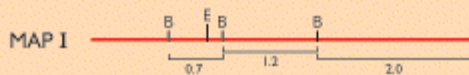
1.0 kb

This must be a BamHI fragment with no internal EcoRI site. We can account for the 1.5 kb EcoRI fragment if we place the 1.0 kb fragment thus:



1.2 kb, 2.0 kb

These must also be BamHI fragments with no internal EcoRI sites. They must lie within the 3.4 kb EcoRI fragment. There are two possibilities:



PREDICTED RESULTS OF A PARTIAL BamHI RESTRICTION

If Map I is correct, then the partial restriction products will include a fragment of $1.2 + 0.7 = 1.9$ kb

If Map II is correct, then the partial restriction products will include a fragment of $2.0 + 0.7 = 2.7$ kb

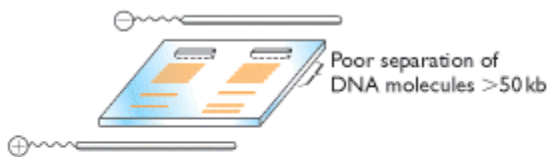


CONCLUSION

Map II is correct

Figure 5.24. Restriction mapping. The objective is to map the *Eco*RI (E) and *Bam*HI (B) sites in a linear DNA molecule of 4.9 kb. The results of single and double restrictions are shown at the top. The sizes of the fragments given after double restriction enable two alternative maps to be constructed, as explained in the central panel, the unresolved issue being the position of one of the three *Bam*HI sites. The two maps are tested by a partial *Bam*HI restriction (bottom), which shows that Map II is the correct one.

(A) Standard agarose gel electrophoresis



(B) Orthogonal field alternation gel electrophoresis (OFAGE)

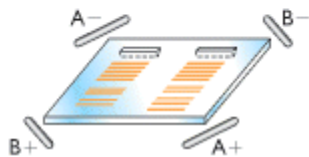


Figure 5.26. Conventional and non-conventional agarose gel electrophoresis. (A) In standard agarose gel electrophoresis the electrodes are placed at either end of the gel and the DNA molecules migrate directly towards the positive electrode. Molecules longer than about 50 kb cannot be separated from one another in this way. (B) In OFAGE, the electrodes are placed at the corners of the gel, with the field pulsing between the A pair and the B pair. OFAGE enables molecules up to 2 Mb to be separated.

Genome mapping

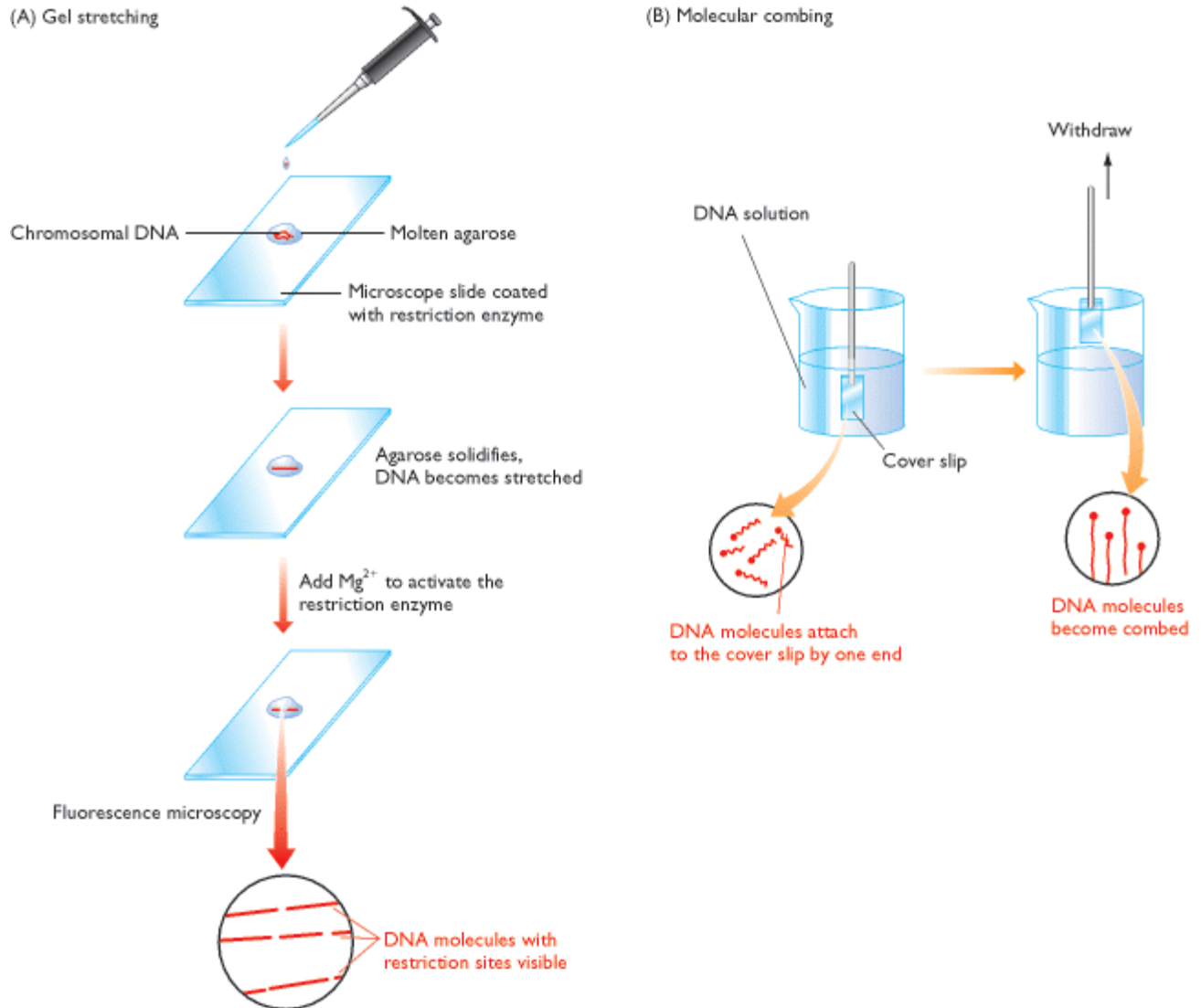


Figure 5.28. Gel stretching and molecular combing. (A) To carry out gel stretching, molten agarose containing chromosomal DNA molecules is pipetted onto a microscope slide coated with a restriction enzyme. As the gel solidifies, the DNA molecules become stretched. It is not understood why this happens but it is thought that fluid movement on the glass surface during gelation might be responsible. Addition of magnesium chloride activates the restriction enzyme, which cuts the DNA molecules. As the molecules gradually coil up, the gaps representing the cut sites become visible. (B) In molecular combing, a cover slip is dipped into a solution of DNA. The DNA molecules attach to the cover slip by their ends, and the slip is withdrawn from the solution at a rate of 0.3 mm s^{-1} , which produces a 'comb' of parallel molecules.

Genome mapping

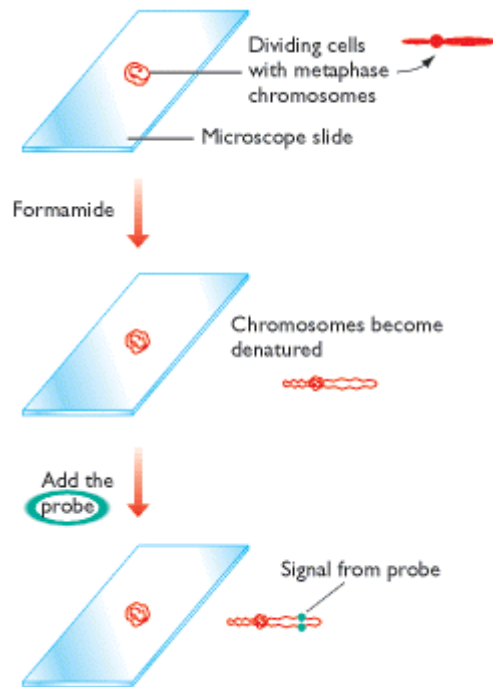


Figure 5.29. Fluorescent *in situ* hybridization. A sample of dividing cells is dried onto a microscope slide and treated with formamide so that the chromosomes become denatured but do not lose their characteristic metaphase morphologies (see [Section 2.2.1](#)). The position at which the probe hybridizes to the chromosomal DNA is visualized by detecting the fluorescent signal emitted by the labeled DNA.

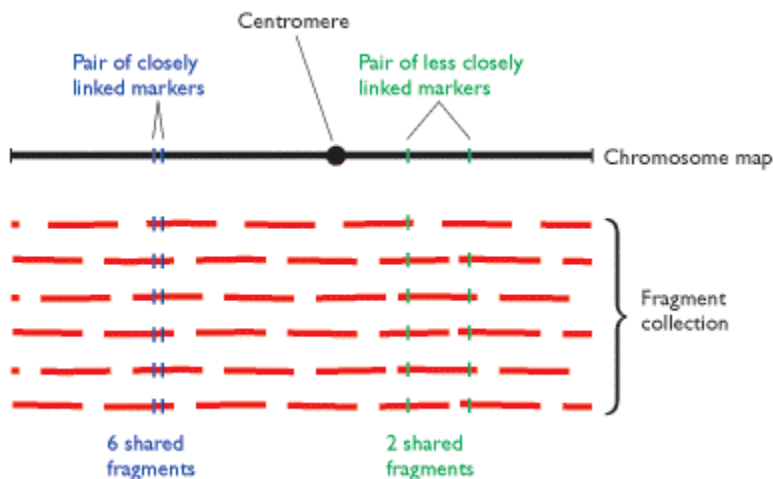
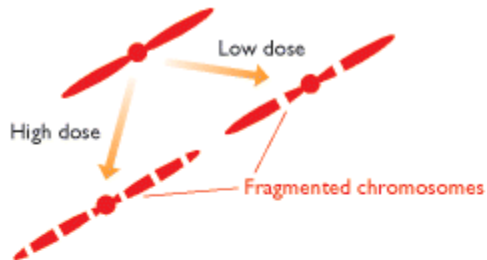


Figure 5.31. A fragment collection suitable for STS mapping. The fragments span the entire length of a chromosome, with each point on the chromosome present in an average of five fragments. The two blue markers are close together on the chromosome map and there is a high

Genome mapping

probability that they will be found on the same fragment. The two green markers are more distant from one another and so are less likely to be found on the same fragment.

(A) Irradiation of chromosomes



(B) Fusion of cells to produce a radiation hybrid

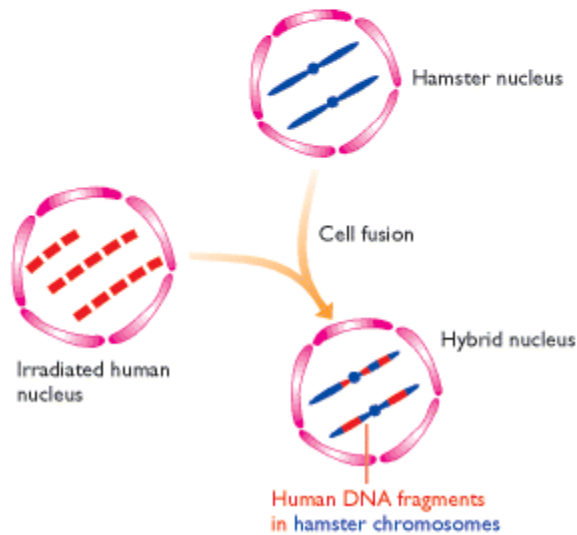


Figure 5.33. Radiation hybrids. (A) The result of irradiation of human cells: the chromosomes break into fragments, smaller fragments generated by higher X-ray doses. In (B), a radiation hybrid is produced by fusing an irradiated human cell with an untreated hamster cell. For clarity, only the nuclei are shown.

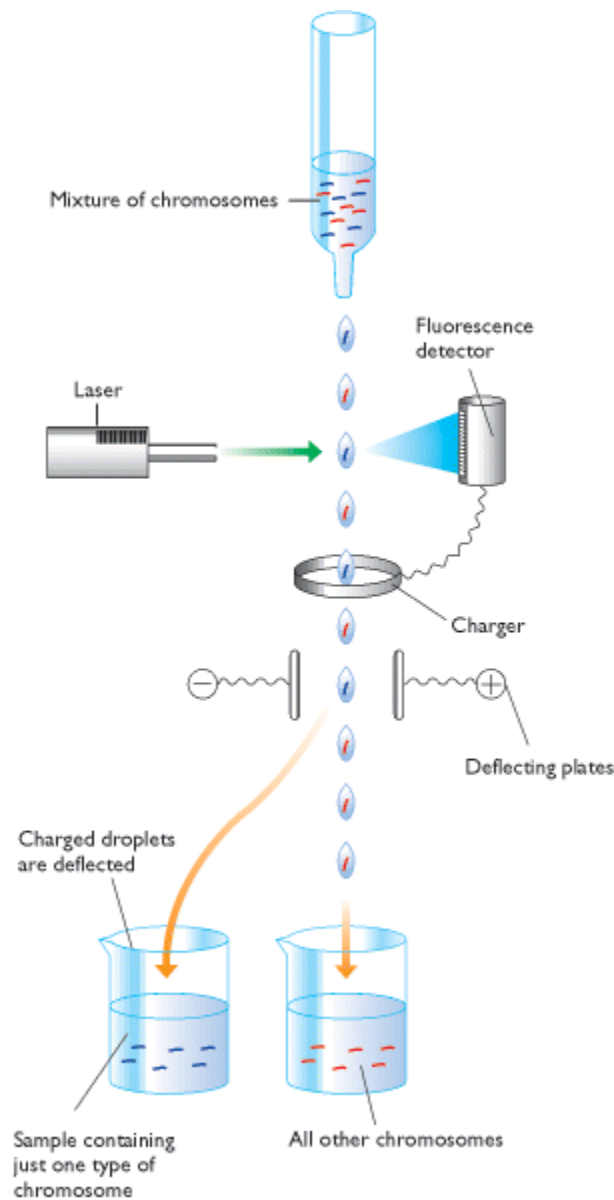


Figure 5.34. Separating chromosomes by flow cytometry. A mixture of fluorescently stained chromosomes is passed through a small aperture so that each drop that emerges contains just one chromosome. The fluorescence detector identifies the signal from drops containing the correct chromosome and applies an electric charge to these drops. When the drops reach the electric plates, the charged ones are deflected into a separate beaker. All other drops fall straight through the deflecting plates and are collected in the waste beaker.

Genome mapping

DNA microarray

A microarray is a multiplex lab-on-a-chip. It is a 2D array on a solid substrate (usually a glass slide or silicon thin-film cell) that assays large amounts of biological material using high-throughput screening miniaturized, multiplexed and parallel processing and detection methods. The concept and methodology of microarrays was first introduced and illustrated in antibody microarrays (also referred to as antibody matrix) by Tse Wen Chang in 1983 in a scientific publication[1] and a series of patents

Types of microarrays include:

DNA microarrays, such as cDNA microarrays, oligonucleotide microarrays, BAC microarrays and SNP microarrays

MMChips, for surveillance of microRNA populations

Protein microarrays

Peptide microarrays, for detailed analyses or optimization of protein-protein interactions

Tissue microarrays

Cellular microarrays (also called transfection microarrays)

Chemical compound microarrays

Antibody microarrays

Carbohydrate arrays (glycoarrays)

Phenotype microarrays

Reverse Phase Protein Microarrays, microarrays of lysates or serum

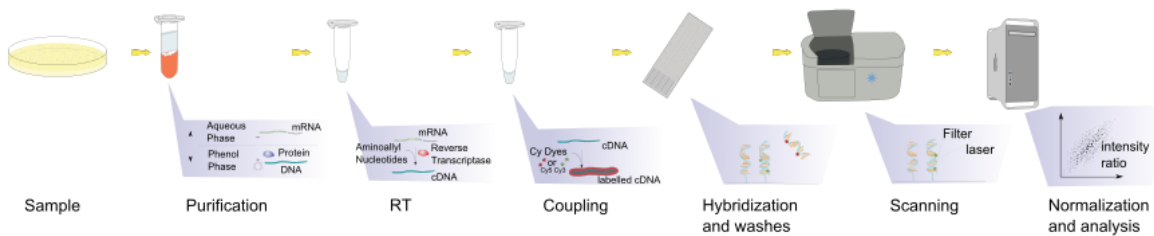
interferometric reflectance imaging sensor (IRIS)

A DNA microarray (also commonly known as DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10⁻¹² moles) of a specific DNA sequence, known as probes (or reporters or oligos). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA (also called anti-sense RNA) sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected

Genome mapping

and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target.

The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. After washing off non-specific bonding sequences, only strongly paired strands will remain hybridized. Fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the hybridization conditions (such as temperature), and washing after hybridization. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantitation in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position.



Uses and types

Many types of arrays exist and the broadest distinction is whether they are spatially arranged on a surface or on coded beads:

The traditional solid-phase array is a collection of orderly microscopic "spots", called features, each with thousands of identical and specific probes attached to a solid surface, such as glass, plastic or silicon biochip (commonly known as a genome chip, DNA chip or gene array). Thousands of these features can be placed in known locations on a single DNA microarray.

The alternative bead array is a collection of microscopic polystyrene beads, each with a specific probe and a ratio of two or more dyes, which do not interfere with the fluorescent dyes used on the target sequence.

Genome mapping

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.

Applications include:

Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.[1]
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms.[2][3]
GeneID	Small microarrays to check IDs of organisms in food and feed (like GMO [1]), mycoplasmas in cell culture, or pathogens for disease detection, mostly combining PCR and microarray technology.
Chromatin immunoprecipitation on Chip	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome. Example protein to immunoprecipitate are histone modifications (H3K27me3, H3K4me2, H3K9me3, etc.), Polycomb-group protein (PRC2:Suz12, PRC1:YY1) and trithorax-group protein (Ash1) to study the epigenetic landscape or RNA Polymerase II to study the transcription landscape.

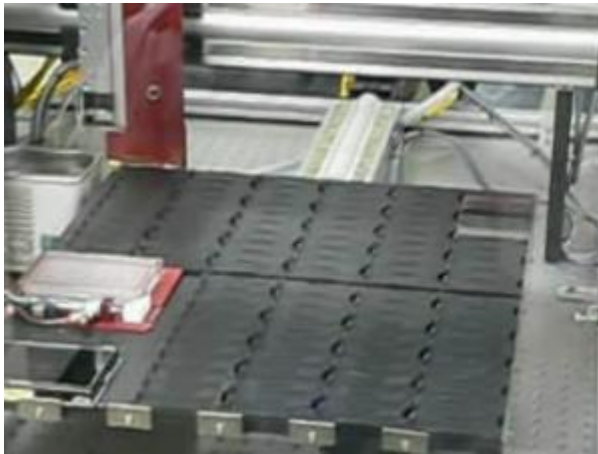
DamID	Analogously to ChIP, genomic regions bound by a protein of interest can be isolated and used to probe a microarray to determine binding site occupancy. Unlike ChIP, DamID does not require antibodies but makes use of adenine methylation near the protein's binding sites to selectively amplify those regions, introduced by expressing minute amounts of protein of interest fused to bacterial DNA adenine methyltransferase.
SNP detection	Identifying single nucleotide polymorphism among alleles within or between populations.[4] Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.
Alternative splicing detection	An exon junction array design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It is of intermediate density, or coverage, to a typical gene expression array (with 1-3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene). It is used to assay the expression of alternative splice forms of a gene. Exon arrays have a different design, employing probes designed to detect each individual exon for known or predicted genes, and can be used for detecting different splicing isoforms.
Fusion genes microarray	A Fusion gene microarray can detect fusion transcripts, e.g. from cancer specimens. The principle behind this is building on the alternative splicing microarrays. The oligo design strategy enables combined measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners.
Tiling array	Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively spliced forms which may not have been

	previously known or predicted.
--	--------------------------------

Fabrication

Microarrays can be manufactured in different ways, depending on the number of probes under examination, costs, customization requirements, and the type of scientific question being asked. Arrays may have as few as 10 probes or up to 2.1 million micrometre-scale probes from commercial vendors.

Spotted vs. in situ synthesised arrays



A DNA microarray being printed by a robot at the University of Delaware

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, photolithography using pre-made masks, photolithography using dynamic micromirror devices, ink-jet printing,[5][6] or electrochemistry on microelectrode arrays.

In spotted microarrays, the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs. The probes are synthesized prior to deposition on the array surface and are then "spotted" onto glass. A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at designated locations on the array surface. The resulting "grid" of probes represents the nucleic acid profiles of the prepared probes and is ready to receive complementary cDNA or cRNA "targets" derived from experimental or clinical samples. This technique is used by research scientists around the world to produce "in-house" printed microarrays from their own labs. These arrays may be easily customized for each experiment,

because researchers can choose the probes and printing locations on the arrays, synthesize the probes in their own lab (or collaborating facility), and spot the arrays. They can then generate their own labeled samples for hybridization, hybridize the samples to the array, and finally scan the arrays with their own equipment. This provides a relatively low-cost microarray that may be customized for each study, and avoids the costs of purchasing often more expensive commercial arrays that may represent vast numbers of genes that are not of interest to the investigator. Publications exist which indicate in-house spotted microarrays may not provide the same level of sensitivity compared to commercial oligonucleotide arrays,[7] possibly owing to the small batch sizes and reduced printing efficiencies when compared to industrial manufactures of oligo arrays. In oligonucleotide microarrays, the probes are short sequences designed to match parts of the sequence of known or predicted open reading frames. Although oligonucleotide probes are often used in "spotted" microarrays, the term "oligonucleotide array" most often refers to a specific technique of manufacturing. Oligonucleotide arrays are produced by printing short oligonucleotide sequences designed to represent a single gene or family of gene splice-variants by synthesizing this sequence directly onto the array surface instead of depositing intact sequences. Sequences may be longer (60-mer probes such as the Agilent design) or shorter (25-mer probes produced by Affymetrix) depending on the desired purpose; longer probes are more specific to individual target genes, shorter probes may be spotted in higher density across the array and are cheaper to manufacture. One technique used to produce oligonucleotide arrays include photolithographic synthesis (Affymetrix) on a silica substrate where light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array.[8] Each applicable probe is selectively "unmasked" prior to bathing the array in a solution of a single nucleotide, then a masking reaction takes place and the next set of probes are unmasked in preparation for a different nucleotide exposure. After many repetitions, the sequences of every probe become fully constructed. More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes.[9]

Two-channel vs. one-channel detection[edit]

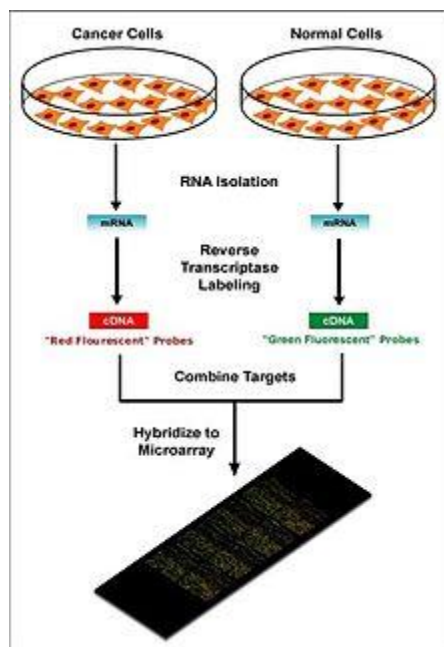


Diagram of typical dual-colour microarray experiment.

Two-color microarrays or two-channel microarrays are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different fluorophores.[10] Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the orange part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.[11]

Oligonucleotide microarrays often carry control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes. Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system. Examples of providers for such microarrays includes Agilent with their Dual-Mode platform, Eppendorf with their DualChip platform for colorimetric Silverquant labeling, and TeleChem International with Arrayit.

In single-channel microarrays or one-color microarrays, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment. Each RNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative.

SBI1309	FUNDAMENTALS OF GENOMICS AND PROTEOMICS	L	T	P	Credits	Total Marks
		3	0	0	3	100

COURSE OBJECTIVES

- The principal objective of this course is for students to acquire knowledge about high throughput tools of genome science.
- This knowledge should enable students to explain the fundamental principles underlying various functional genomics techniques and their applications in various biological systems.

UNIT 1 INTRODUCTION TO GENOMICS**9 Hrs.**

Organization and main features of prokaryotic and eukaryotic genomes. Genome sequencing methods: Maxim - Gilbert and Sanger's method, automated sequencing, pyro-sequencing. Sequence assembly: Clone contig and shotgun approaches. Genome Projects: Aims and objectives, Brief outlook of various Genome Projects - Human genome Project, Plant and animal genome projects. Genome databases

UNIT 2 GENOME MAPPING**9 Hrs.**

Genetic and physical maps. RFLP, SSLP, STRs, VNTRs, EST, STS, FISH, Radiation hybrids. Sequence markers - SNPs. Determination of the functions of genes: gene inactivation (knock-out, anti-sense and RNA interference) and gene over expression. Gene expression analysis - DNA microarray approach.

UNIT 3 INTRODUCTION TO PROTEOMICS**9 Hrs.**

Introduction, Branches of proteomics. Proteome Project. Interactions in Proteomics: Protein-Protein Interactions. Methods - Yeast Two hybrid analysis, Phage display, Databases. Protein-DNA Interactions - DNA binding Motifs.

UNIT 4 PROTEOMIC STUDIES**9 Hrs.**

Characterization of Proteome: Protein purification, Protein separation: 2-D gel electrophoresis and affinity chromatography, HPLC. Protein identification: Mass Spectrometry. Protein sequence analysis: N-terminal determination methods. Protein modifications. Protein expression profiling: Protein microarrays.

UNIT 5 OMICS CONCEPTS**9 Hrs.**

Transcriptomics and SAGE; Pharmacogenomics, Comparative Genomics, Population genomics, Metabolomics & KEGG, Fluxomics, Glycomics, Nutrigenomics, Epigenomics

Determining the Function of Genes during Development

Transgenic cells and organisms

While it is important to know the sequence of a gene and its temporal-spatial pattern of expression, what's really crucial is to know the functions of that gene during development. Recently developed techniques have enabled us to study gene function by moving certain genes into and out of embryonic cells.

WEBSITE

4.6 Bioinformatics. Information about gene regulation and developmental pathways may soon be modeled on computers. Accessibility to this information may enable researchers to design experiments that have higher chances of success. <http://www.devbio.com/chap04/link0406.shtml>

Inserting new DNA into a cell

Cloned pieces of DNA can be isolated, modified (if so desired), and placed into cells by several means. One very direct technique is **microinjection**, in which a solution containing the cloned gene is injected very carefully into the nucleus of a cell (Capecchi 1980). This is an especially useful technique for injecting genes into newly fertilized eggs, since the haploid nuclei of the sperm and egg are relatively large (Figure 4.18). In **transfection**, DNA is incorporated directly into cells by incubating them in a solution that makes them “drink” it in. The chances of a DNA fragment being incorporated into the chromosomes in this way are relatively small, however, so the DNA of interest is usually mixed with another gene, such as a gene encoding resistance to a particular antibiotic, that enables those rare cells that incorporate the DNA to survive under culture conditions that will kill all the other cells (Perucho et al. 1980; Robins et al. 1981). Another technique is **electroporation**, in which a high-voltage pulse “pushes” the DNA into the cells.



Figure 4.18

Insertion of new DNA into embryonic cells. Here, DNA (from cloned genes) is injected into the pronucleus of a mouse egg. (From Wagner et al. 1981; photograph courtesy of T. E. Wagner.)

A more “natural” way of getting genes into cells is to put the cloned gene into a **transposable element** or **retroviral vector**. These are naturally occurring mobile regions of DNA that can integrate themselves into the genome of an organism. Retroviruses are RNA-containing viruses. Within a host cell, they make a DNA copy of themselves (using their own virally encoded reverse transcriptase); the copy then becomes double-stranded and integrates itself into a host chromosome. The integration is accomplished by two identical sequences (long terminal repeats) at the ends of the retroviral DNA. Retroviral vectors are made by removing the viral packaging genes (needed for the exit of viruses from the cell) from the center of a mouse retrovirus. This extraction creates a vacant site where other genes can be placed. By using the appropriate restriction enzymes, researchers can insert an isolated gene (such as a gene isolated by PCR) and insert it into a retroviral vector. These retroviral vectors infect mouse cells with an efficiency approaching 100%. Similarly, in *Drosophila*, new genes can be carried into a fly via **P elements**. These DNA sequences are naturally occurring transposable elements that can integrate like viruses into any region of the *Drosophila* genome. Moreover, they can be isolated, and cloned genes can be inserted into the center of the P element. When the recombined P element is injected into a *Drosophila* oocyte, it can integrate into the DNA and provide the embryo with the new gene (Spradling and Rubin 1982).

Chimeric mice

The techniques described above have been used to transfer genes into every cell of the mouse embryo (Figure 4.19). During early mouse development, there is a stage when only two cell types are present: the outer trophoblast cells, which will form the fetal portion of the placenta, and the inner cell mass, whose cells will give rise to the embryo itself. These inner cells are the cells whose separation can lead to twins (Chapters 3 and 11), and if an inner cell mass blastomere of one mouse is transferred into the embryo of a second mouse, that donor cell can contribute to every organ of the host embryo.

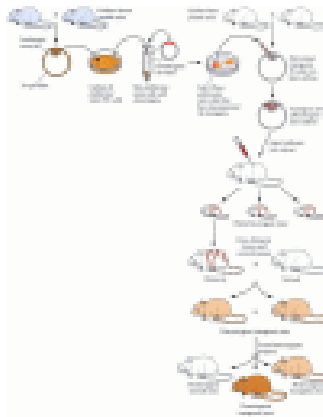


Figure 4.19

Production of transgenic mice. Embryonic stem cells from one mouse are cultured and their genome altered by the addition of a cloned gene. These transgenic cells are selected and then injected into the early stages of a host mouse embryo. Here, the transgenic (more...)

Inner cell mass blastomeres can be isolated from the embryo and cultured in vitro; such cultured cells are called **embryonic stem cells (ES cells)**. ES cells retain their totipotency, and each of them can contribute to all organs if injected into a host embryo (Gardner 1968; Moustafa and Brinster 1972). Moreover, once in culture, these cells can be treated as described in the preceding section so that they will incorporate new DNA. A treated ES cell (the entire cell, not just the DNA) can then be injected into another early-stage mouse embryo, and will integrate into the host embryo. The result is a **chimeric mouse**.^{*} Some of this mouse's cells will be derived from the host embryonic stem cells, but some portion of its cells will be derived from the treated embryonic stem cell. If the treated cells become part of the germ line of the mouse, some of its gametes will be derived from the donor cell. If such a chimeric mouse is mated with a wild-type mouse, some of its progeny will carry one copy of the inserted gene. When these heterozygous progeny are mated to one another, about 25% of the resulting offspring will carry two copies of the inserted gene in every cell of their bodies (Gossler et al. 1986). Thus, in three generations—the chimeric mouse, the heterozygous mouse, and the homozygous mouse—a gene that was cloned from some other organism will be present in both copies of the chromosomes within the mouse genome. Strains of such transgenic mice have been particularly useful in determining how genes are regulated during development.

Gene targeting (“knockout”) experiments

The analysis of early mammalian embryos has long been hindered by our inability to breed and select animals with mutations that affect early embryonic development. This block has been circumvented by the techniques of **gene targeting** (or, as it is sometimes called, **gene knockout**). These techniques are similar to those that generate transgenic mice, but instead of *adding* genes, gene targeting *replaces* wild-type alleles with mutant ones. As an example, we will look at the gene knockout of bone morphogenetic protein 7 (BMP7). Bone morphogenetic proteins are involved in numerous developmental interactions whereby one set of cells interacts with other neighboring cells to alter their properties. BMP7 has been implicated as a protein that prevents cell death and promotes cell division in several developing organs. Dudley and his colleagues 1995 used gene targeting to find the function of BMP7 in the development of the mouse. First, they isolated the *BMP7* gene, cut it at one site with a restriction enzyme, and inserted a bacterial gene for neomycin resistance into that site (Figure 4.20). In other words, they mutated the *BMP7* gene by inserting into it a large piece of foreign DNA, destroying the ability of the BMP7 protein to function. These mutant *BMP7* genes were electroporated into ES cells that were sensitive to neomycin. Once inside the nucleus of an ES cell, the mutated *BMP7* gene may replace a normal allele of *BMP7* by a process called homologous recombination. In this process, the enzymes involved in DNA repair and replication incorporate the mutant gene in the place of the normal copy. It's a rare event, but such cells can be selected by growing the ES cells in neomycin. Most of the cells are killed by the drug, but the ones that have acquired resistance from the incorporated gene survive. The resulting cells have one normal *BMP7* gene and one mutated *BMP7* gene. These heterozygous ES cells were then microinjected into mouse blastocysts, where they were integrated into the cells of the embryo. The resulting mice were chimeras composed of wild-type cells from the host embryo and heterozygous *BMP7*-containing cells from the donor ES cells. The chimeras were mated to wild-type mice, producing progeny that were heterozygous for the *BMP7* gene. These heterozygous mice were then bred with each other, and about 25% of their progeny carried two copies of the mutated *BMP7* gene. These homozygous mutant mice lacked eyes and kidneys (Figure 4.21). In the absence of BMP7, it appears that many of the cells that normally form these two organs stop dividing and die. In this way, gene targeting can be used to analyze the roles of particular genes during mammalian development.

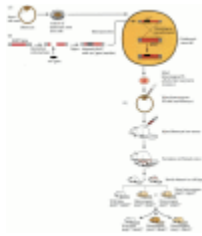


Figure 4.20

Technique for gene targeting. In this case, the targeted gene is *BMP7*. (A) Embryonic stem (ES) cells from a mouse blastocyst are cultured. (B) Cloned *BMP7* genes are cut with a restriction enzyme, and a neomycin resistance gene is inserted into the region (more...)



Figure 4.21

Morphological analysis of *BMP7* knockout mice. A wild-type (A) and a homozygous *BMP7*-deficient mouse (B) at day 17 of their 21-day gestation. The *BMP7*-deficient mouse lacks eyes. The kidneys of these mice at day 19 of gestation are shown in (C). The kidney (more...)

Go to:

Determining the function of a message: Antisense RNA

Another method for determining the function of a gene is to use “antisense” copies of its message to block the function of that message. Antisense RNA allows developmental biologists to determine the function of genes during development and to analyze the action of genes that would otherwise be inaccessible for genetic analysis.

Antisense messages can be generated by cloning DNA into vectors that have promoters at both ends of the inserted gene. When incubated with a particular RNA polymerase and nucleotide triphosphates, the promoter will initiate transcription of the message “in the wrong direction.” In so doing, it synthesizes a transcript that is complementary to the natural one (Figure 4.23A). This complementary transcript is called **antisense RNA** because it is the complement of the original (“sense”) message. When large amounts of antisense RNA are injected or transfected into cells containing the normal mRNA from the same gene, the antisense RNA binds to the normal message, and the resulting double-stranded nucleic acid is degraded. (Cells have enzymes to digest double-stranded nucleic acids in the cytoplasm.) This causes a functional deletion of the message, just as if there were a deletion mutation for that gene.

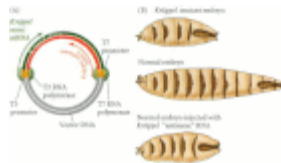


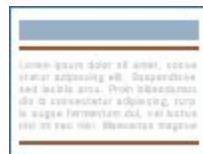
Figure 4.23

Use of antisense RNA to examine the roles of genes in development. (A) An antisense message (in this case, to the *Krüppel* gene of *Drosophila*) is produced by placing the cloned cDNA fragment for the *Krüppel* message between two strong promoters (more...)

The similarities between the phenotypes produced by a loss-of-function mutation and by antisense RNA treatment were seen when antisense RNA was made to the *Krüppel* gene of *Drosophila*. *Krüppel* is critical for forming the thorax and abdomen of the fly. If this gene is absent, fly larvae die because they lack thoracic and anterior abdominal segments (Figure 4.23B). A similar situation is created when large amounts of antisense RNA against the *Krüppel* message are injected into early fly embryos (Rosenberg et al. 1985).

WEBSITE

4.8 RNA interference. Soaking the nematode *C. elegans* in a solution containing double-stranded RNA will knock out the expression of that gene not only in the soaked animal, but also in the progeny of that animal. <http://www.devbio.com/chap04/link0408.shtml>



Box

Human Somatic and Germ Line Gene Therapy. *Embryonic Stem Cells* In 1998, two laboratories (Gearhart 1998; Thomson et al. 1998) announced that they had derived human embryonic stem cells. In some instances, these cells were derived from inner cell masses (more...)

Go to:

Footnotes

*

It is critical to note the difference between a chimera and a hybrid. A **hybrid** results from the union of two different genomes within the same cell: the offspring of an AA genotype parent and an aa genotype parent is an Aa hybrid. A **chimera** results when cells of

different genetic constitution appear in the same organism. The term is apt: it refers to a mythical beast with a lion's head, a goat's body, and a serpent's tail.

Gene targeting (also, replacement strategy based on homologous recombination) is a genetic technique that uses homologous recombination to change an endogenous gene. The method can be used to delete a gene, remove exons, add a gene, and introduce point mutations. Gene targeting can be permanent or conditional. Conditions can be a specific time during development / life of the organism or limitation to a specific tissue, for example. Gene targeting requires the creation of a specific vector for each gene of interest. However, it can be used for any gene, regardless of transcriptional activity or gene size.

Methods[edit]

Gene targeting methods are established for several model organisms and may vary depending on the species used. In general, a targeting construct made out of DNA is generated in bacteria. It typically contains part of the gene to be targeted, a reporter gene, and a (dominant) selectable marker.

To target genes in mice, this construct is then inserted into mouse embryonic stem cells in culture. After cells with the correct insertion have been selected, they can be used to contribute to a mouse's tissue via embryo injection. Finally, chimeric mice where the modified cells made up the reproductive organs are selected for via breeding. After this step the entire body of the mouse is based on the previously selected embryonic stem cell.

To target genes in moss, this construct is incubated together with freshly isolated protoplasts and with Polyethylene glycol. As mosses are haploid organisms,^[2] regenerating moss filaments (protonema) can directly be screened for gene targeting, either by treatment with antibiotics or with PCR. Unique among plants, this procedure for reverse genetics is as efficient as in yeast.^[3] Using modified procedures, gene targeting has also been successfully applied to cattle, sheep, swine, and many fungi.

The frequency of gene targeting can be significantly enhanced through the use of engineered endonucleases such as zinc finger nucleases,^[4] engineered homing endonucleases,^[5] and nucleases based on engineered TAL effectors.^[6] To date, this method has been applied to a number of species including *Drosophila melanogaster*,^[4] tobacco,^{[7][8]} corn,^[9] human cells,^[10] mice,^[11] and rats.^[11]

Comparison with gene trapping

Gene trapping is based on random insertion of a cassette while gene targeting targets a specific gene. Cassettes can be used for many different things while the flanking homology regions of gene targeting cassettes need to be adapted for each gene. This makes gene trapping more easily amenable for large scale projects than targeting. On the other hand, gene targeting can be used for genes with low transcriptions that would go undetected in a trap screen. Also, the probability of trapping increases with intron size. For gene targeting these compact genes are just as easily altered.

Applications

Gene targeting has been widely used to study human genetic diseases by removing ("knocking out"), or adding ("knocking in"), specific mutations of interest to a variety of models. Previously used to engineer rat cell models, advances in gene targeting technologies are enabling the creation of a new wave of isogenic human disease models. These models are the most accurate in-vitro models available to researchers to date, and are facilitating the development of new personalized drugs and diagnostics, particularly in the field of cancer.^[12]

2007 Nobel prize

Mario R. Capecchi, Martin J. Evans and Oliver Smithies were declared laureates of the 2007 Nobel Prize in Physiology or Medicine for their work on "principles for introducing specific gene modifications in mice by the use of embryonic stem cells", or gene targeting

Gene knock out

A **gene knockout** (abbreviation: **KO**) is a genetic technique in which one of an organism's genes is made inoperative ("knocked out" of the organism). Also known as **knockout organisms** or simply **knockouts**, they are used in learning about a gene that has been sequenced, but which has an unknown or incompletely known function. Researchers draw inferences from the difference between the knockout organism and normal individuals.

The term also refers to the process of creating such an organism, as in "knocking out" a gene. The technique is essentially the opposite of a gene knockin. Knocking out two genes simultaneously in an organism is known as a **double knockout (DKO)**. Similarly the

terms **triple knockout (TKO)** and **quadruple knockouts (QKO)** are used to describe three or four knocked out genes, respectively.

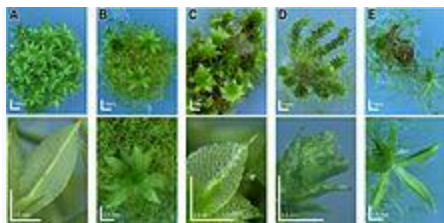
Method



A laboratory mouse in which a gene affecting hair growth has been knocked out (left), is shown next to a normal lab mouse.

Knockout is accomplished through a combination of techniques, beginning in the test tube with a plasmid, a bacterial artificial chromosome or other DNA construct, and proceeding to cell culture. Individual cells are genetically transfected with the DNA construct. Often the goal is to create a transgenic animal that has the altered gene. If so, embryonic stem cells are genetically transformed and inserted into early embryos. Resulting animals with the genetic change in their germline cells can then often pass the gene knockout to future generations.

To create knockout moss, transfection of protoplasts is the preferred method. Such transformed *Physcomitrella*-protoplasts directly regenerate into fertile moss plants. Eight weeks after transfection, the plants can be screened for gene targeting via PCR.^[1]



Wild-type Physcomitrella and knockout mosses: Deviating phenotypes induced in gene-disruption library transformants. *Physcomitrella* wild-type and transformed plants were grown on minimal Knop medium to induce differentiation and development of gametophores. For each plant, an overview (upper row; scale bar corresponds to 1 mm) and a close-up (bottom row; scale bar equals 0.5 mm) are shown. A: Haploid wild-type moss plant completely covered with leafy gametophores and close-up of wild-type leaf. B–D: Different mutants.^[2]

The construct is engineered to recombine with the target gene, which is accomplished by incorporating sequences from the gene itself into the construct. Recombination then occurs in the region of that sequence within the gene, resulting in the insertion of a foreign sequence to disrupt the gene. With its sequence interrupted, the altered gene in most cases will be translated into a nonfunctional protein, if it is translated at all.



A knockout mouse (left) that is a model of obesity, compared with a normal mouse.

A conditional knockout allows gene deletion in a tissue or time specific manner. This is done by introducing short sequences called loxP sites around the gene. These sequences will be introduced into the germ-line via the same mechanism as a knock-out. This germ-line can then be crossed to another germline containing Cre-recombinase which is a viral enzyme that can recognize these sequences, recombines them and deletes the gene flanked by these sites.

Because the desired type of DNA recombination is a rare event in the case of most cells and most constructs, the foreign sequence chosen for insertion usually includes a reporter. This enables easy selection of cells or individuals in which knockout was successful. Sometimes the DNA construct inserts into a chromosome without the desired homologous recombination with the target gene. To eliminate such cells, the DNA construct often contains a second region of DNA that allows such cells to be identified and discarded.

In diploid organisms, which contain two alleles for most genes, and may as well contain several related genes that collaborate in the same role, additional rounds of transformation and selection are performed until every targeted gene is knocked out. Selective breeding may be required to produce homozygous knockout animals.

Gene knockin is similar to gene knockout, but it replaces a gene with another instead of deleting it.

Use

Knockouts are primarily used to understand the role of a specific gene or DNA region by comparing the knockout organism to a wildtype with a similar genetic background.

Genome mapping

Knockouts organisms are also used as screening tools in the development of drugs, to target specific biological processes or deficiencies by using a specific knockout, or to understand the mechanism of action of a drug by using a library of knockout organisms spanning the entire genome, such as in *Saccharomyces cerevisiae*

Gene knockout

A gene knockout is a genetically engineered organism that carries one or more genes in its chromosomes that have been made inoperative (have been "knocked out" of the organism). This is done for research purposes. Also known as knockout organisms or simply knockouts, they are used in learning about a gene that has been sequenced, but which has an unknown or incompletely known function. Researchers draw inferences from the difference between the knockout organism and normal individuals. The term also refers to the process of creating such an organism, as in "knocking out" a gene. Knockout is accomplished through a combination of techniques, beginning in the test tube with a plasmid, a bacterial artificial chromosome or other DNA construct, and proceeding to cell culture. Individual cells are genetically transformed with the construct and--for knockouts in multi-cellular organisms--ultimately fused with a stem cell from a nascent embryo. The construct is engineered to recombine with the target gene, which is accomplished by incorporating sequences from the gene itself into the construct. Recombination then occurs in the region of that sequence within the gene, resulting in the insertion of a foreign sequence to disrupt the gene. With its sequence interrupted, the altered gene in most cases will be translated into a nonfunctional protein, if it is translated at all. A conditional knockout allows gene deletion in a tissue specific manner. Because recombination is a rare event in the case of most cells and most constructs, the foreign sequence chosen for insertion usually is a reporter. This enables easy selection of cells or individuals in which knockout was successful. In diploid organisms, which contain two alleles for most genes, and may as well contain several related genes that collaborate in the same role, additional rounds of transformation and selection are performed until every targeted gene is knocked out. Knock-in is similar to knock-out, but instead it replaces a gene with another instead of deleting it. Knockout mouse A knockout mouse is a genetically engineered mouse one or more of whose genes have been made inoperable through a gene knockout. Knockout is a route to learning about a gene that has been sequenced but has an unknown or incompletely known function. Mice are the laboratory animal species most closely related to humans in which the knockout technique can be easily performed, so they are a favourite subject for knockout experiments, especially with regard to genetic questions that relate to human physiology. (Gene knockout in rats is much harder and has only been possible since 2003.) Use Knocking out the activity of a gene provides information about what that gene

normally does. Humans share many genes with mice. Consequently, observing the characteristics of knockout mice gives researchers information that can be used to better understand how a similar gene may cause or contribute to disease in humans. Examples of research in which knockout mice have been useful include studying and modelling different kinds of cancer, obesity, heart disease, diabetes, arthritis, substance abuse, anxiety, aging and Parkinson disease. Knockout mice also offer a biological context in which drugs and other therapies can be developed and tested. Many of these mouse models are named after the gene that has been inactivated. For example, the p53 knockout mouse is named after the p53 gene which codes for a protein that normally suppresses the growth of tumours by arresting cell division. Humans born with mutations that inactivate the p53 gene suffer from Li-Fraumeni syndrome, a condition that dramatically increases the risk of developing bone cancers, breast cancer and blood cancers at an early age. Other mouse models are named, often with creative flair, according to their physical characteristics or behaviours. For example, "Methuselah" is a knockout mouse model noted for longevity, while "Frantic" is a model useful for studying anxiety disorders.

Procedure There are several variations to the procedure of producing knockout mice; the following is a typical example.

1. The gene to be knocked out is isolated from a mouse gene library. Then a new DNA sequence is engineered which is very similar to the original gene and its immediate neighbor sequence, except that it is changed sufficiently to make it inoperable. Usually, the new sequence is also given a marker gene, a gene that normal mice don't have and that transfers resistance to a certain antibiotic or a selectable marker.
2. From a mouse morula (a very young embryo consisting of a ball of undifferentiated cells), stem cells are isolated; these can be grown in vitro. For this example, we will take a stem cell from a white mouse.
3. The stem cells from step 2 are combined with the new sequence from step 1. This is done via electroporation (using electricity to transfer the DNA across the cell membrane). Some of the electroporated stem cells will incorporate the new sequence into their chromosomes in place of the old gene; this is called homologous recombination. The reason for this process is that the new and the old sequence are very similar. Using the antibiotic from step 1, those stem cells that actually did incorporate the new sequence can be quickly isolated from those that did not.
4. The stem cells from step 3 are inserted into mouse blastocyst cells. For this example, we use blastocysts from a grey mouse. These blastocysts are then implanted into the uterus of female mice, to complete the pregnancy. The blastocysts contain two types of stem cells: the original ones (grey mouse), and the newly

engineered ones (white mouse). The newborn mice will therefore be chimeras: parts of their bodies result from the original stem cells, other parts result from the engineered stem cells. Their furs will show patches of white and grey. 5. Newborn mice are only useful if the newly engineered sequence was incorporated into the germ cells (egg or sperm cells). So we cross these new mice with others and watch for offspring that are all white. These are then further inbred to produce mice that carry no functional copy of the original gene.

Limitations

While knockout mice technology represents a valuable research tool, some important limitations exist. About 15 percent of gene knockouts are developmentally lethal, which means that the genetically altered embryos cannot grow into adult mice. The lack of adult mice limits studies to embryonic development and often makes it more difficult to determine a gene's function in relation to human health. In some instances, the gene may serve a different function in adults than in developing embryos. Knocking out a gene also may fail to produce an observable change in a mouse or may even produce different characteristics from those observed in humans in which the same gene is inactivated. For example, mutations in the p53 gene are associated with more than half of human cancers and often lead to tumours in a particular set of tissues. However, when the p53 gene is knocked out in mice, the animals develop tumours in a different array of tissues. There is variability in the whole procedure depending largely on the strain from which the stem cells have been derived. Generally cells derived from strain 129 are used. This specific strain is not suitable for many experiments (e.g., behavioural), so it is very common to backcross the offspring to other strains. Some genomic loci have been proven very difficult to knock out. Reasons might be the presence of repetitive sequences, extensive DNA methylation, or heterochromatin.

Antisense RNA technology

Antisense RNA (asRNA) is a single-stranded RNA that is complementary to a messenger RNA (mRNA) strand transcribed within a cell. Some authors have used the term micRNA (mRNA-interfering complementary RNA) to refer to these RNAs but it is not widely used.^[1]

Antisense RNA may be introduced into a cell to inhibit translation of a complementary mRNA by base pairing to it and physically obstructing the translation machinery.^[2] This effect is therefore stoichiometric. An example of naturally occurring mRNA antisense mechanism is the *hok/sok* system of the *E. coli* R1 plasmid. Antisense RNA has long been thought of as a promising technique for disease therapy; the first antisense therapeutic to reach the market is the drug fomivirsen, approved in 1998. Mipomersen was approved in the United States in 2013. One commentator has characterized antisense RNA as one of "dozens of technologies that are gorgeous in concept, but exasperating in [commercialization]".^[3] Generally, antisense RNA still lack effective design, biological activity, and efficient route of administration.^[4]

The effects of antisense RNA are related with the effects of RNA interference (RNAi). The RNAi process, found only in eukaryotes, is initiated by double-stranded RNA fragments, which may be created by the expression of an anti-sense RNA followed by the base-pairing of the anti-sense strand to the target transcript.^[5] Double-stranded RNA may be created by other mechanisms (including secondary RNA structure). The double-stranded RNA is cleaved into small fragments by DICER, and then a single strand of the fragment is incorporated into the RNA-induced silencing complex (RISC) so that the RISC may bind to and degrade the complementary mRNA target.^[6] Some genetically engineered transgenic plants that express antisense RNA do activate the RNAi pathway.^[7] This processes resulted in differing magnitudes of gene silencing induced by the expression of antisense RNA. Well-known examples include the FlavrSavr tomato and two cultivars of ringspot-resistant papaya.^{[8][9]}

Transcription of longer *cis*-antisense transcripts is a common phenomenon in the mammalian transcriptome.^[10] Although the function of some cases have been described, such as the Zeb2/Sip1 antisense RNA, no general function has been elucidated. In the case of Zeb2/Sip1,^[11] the antisense noncoding RNA is opposite the 5' splice site of an intron in the 5'UTR of the Zeb2 mRNA. Expression of the antisense ncRNA prevents splicing of an intron that contains a ribosome entry site necessary for efficient expression of the Zeb2 protein.

Genome mapping

Transcription of long antisense ncRNAs is often concordant with the associated protein-coding gene,^[12] but more detailed studies have revealed that the relative expression patterns of the mRNA and antisense ncRNA are complex



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

UNIT – III - Fundamentals of Genomics and Proteomics– SBI1309

PROTEOMICS

Proteomics is the large-scale study of proteins, particularly their structures and functions.^{[1][2]} Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term "proteomics" was coined to make an analogy with genomics, the study of the genes. The word "proteome" is a portmanteau of "*protein*" and "*genome*". The proteome of an organism is the set of proteins produced by it during its life, and its genome is its set of genes.

Proteomics is often considered the next step in the study of biological systems, after genomics. It is much more complicated than genomics, mostly because while an organism's genome is rather constant, a proteome differs from cell to cell and constantly changes through its biochemical interactions with the genome and the environment. One organism has radically different protein expression in different parts of its body, different stages of its life cycle and different environmental conditions. Another major difficulty is the complexity of proteins relative to nucleic acids. E.g., in human there are about 25 000 identified genes but an estimated >500 000 proteins that are derived from these genes. This increased complexity derives from mechanisms such as alternative splicing, protein modification (glycosylation, phosphorylation) and protein degradation.

Scientists are very interested in proteomics because it gives a much better understanding of an organism than genomics. First, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. Second, many proteins experience post-translational modifications that profoundly affect their activities; for example some proteins are not active until they become phosphorylated. Methods such as phosphoproteomics and glycoproteomics are used to study post-translational modifications. Third, many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications. Finally, many proteins form complexes with other proteins or RNA molecules, and only function in the presence of these other molecules.

Since proteins play a central role in the life of an organism, proteomics is instrumental in discovery of biomarkers, such as markers that indicate a particular disease.

With the completion of a rough draft of the human genome, many researchers are looking at how genes and proteins interact to form other proteins. A surprising finding of the Human Genome Project is that there are far fewer protein-coding genes in the human genome

than proteins in the human proteome (20,000 to 25,000 genes vs. > 500,000 proteins). The human body may even contain more than 2 million proteins, each having different functions. The protein diversity is thought to be due to alternative splicing and post-translational modification of proteins. The discrepancy implies that protein diversity cannot be fully characterized by gene expression analysis, thus proteomics is useful for characterizing cells and tissues.

To catalog all human proteins, their functions and interactions is a great challenge for scientists. An international collaboration with these goals is co-ordinated by the Human Proteome Organization (HUPO).

The Proteome

The proteome is the protein complement of the genome. It is quite a bit more complicated than the genome because a single gene can give rise to a number of different proteins through

- alternative splicing of the pre-messenger RNAs;
- RNA editing of the pre-messenger RNAs;
- attachment of carbohydrate residues to form glycoproteins);
- addition of phosphate groups to some of the amino acids in the protein [Examples];
- etc.

While we humans may turn out to have only 25 to 30 thousand genes, we probably make at least 10 times that number of different proteins. More than 50% of our genes produce pre-mRNAs that are alternatively-spliced.

The study of proteomics is important because proteins are responsible for both the structure and the functions of all living things. Genes are simply the instructions for making proteins. It is proteins that make life.

The set of proteins within a cell varies

- from one differentiated cell type to another (e.g. red blood cell vs lymphocyte) and
- from moment to moment, depending on the activities of the cell, e.g.,
 - getting ready to duplicate its genome;
 - repairing damage to its DNA;
 - responding to a newly-available nutrient [Example] or cytokine;
 - responding to the arrival of a hormone

Studying proteomics

Most proteins function in collaboration with other proteins, and one goal of proteomics is to identify which proteins interact. This often gives important clues about the functions of newly discovered proteins. Several methods are available to probe protein-

protein interactions. The traditional method is yeast two-hybrid analysis. New methods include protein microarrays, immunoaffinity chromatography followed by mass spectrometry, and combinations of experimental methods such as phage display and computational methods.

Current research in proteomics requires first that proteins be resolved, sometimes on a massive scale. Protein separation can be performed using two-dimensional gel electrophoresis, which usually separates proteins first by isoelectric point and then by molecular weight. Protein spots in a gel can be visualized using a variety of chemical stains or fluorescent markers. Proteins can often be quantified by the intensity of their stain. Once proteins are separated and quantified, they are identified. Individual spots are cut out of the gel and cleaved into peptides with proteolytic enzymes. These peptides can then be identified by mass spectrometry, specifically matrix-assisted laser desorption-ionization time-of-flight (MALDI-TOF) mass spectrometry. In this procedure, a peptide is placed on a matrix, which causes the peptide to form crystals^[citation needed]. Then the peptide on the matrix is ionized with a laser beam and an increase in voltage at the matrix is used to shoot the ions toward a detector in which the time it takes an ion to reach the detector depends on its mass. The higher the mass, the longer the time of flight of the ion. In a MALDI-TOF mass spectrometer, the ions can also be deflected with an electrostatic reflector that also focuses the ion beam. Thus, the masses of the ions reaching the second detector can be determined with high precision and these masses can reveal the exact chemical compositions of the peptides, and therefore their identities.

Protein mixtures can also be analyzed without prior separation. These procedures begin with proteolytic digestion of the proteins in a complex mixture. The resulting peptides are often injected onto a high pressure liquid chromatography column (HPLC) that separates peptides based on hydrophobicity. HPLC can be coupled directly to a time-of-flight mass spectrometer using electrospray ionization. Peptides eluting from the column can be identified by tandem mass spectrometry (MS/MS). The first stage of tandem MS/MS isolates individual peptide ions, and the second breaks the peptides into fragments and uses the fragmentation pattern to determine their amino acid sequences. Labeling with isotope tags can be used to quantitatively compare proteins concentration among two or more protein samples.

One of the most promising developments to come from the study of human genes and proteins has been the identification of potential new drugs for the treatment of disease. This relies on genome and proteome information to identify proteins associated with a disease,

which computer software can then use as targets for new drugs. For example, if a certain protein is implicated in a disease, its 3D structure provides the information to design drugs to interfere with the action of the protein. A molecule that fits the active site of an enzyme, but cannot be released by the enzyme, will inactivate the enzyme. This is the basis of new drug-discovery tools, which aim to find new drugs to inactivate proteins involved in disease. As genetic differences among individuals are found, researchers expect to use these techniques to develop personalized drugs that are more effective for the individual.

A computer technique which attempts to fit millions of small molecules to the three-dimensional structure of a protein is called "virtual ligand screening". The computer rates the quality of the fit to various sites in the protein, with the goal of either enhancing or disabling the function of the protein, depending on its function in the cell. A good example of this is the identification of new drugs to target and inactivate the HIV-1 protease. The HIV-1 protease is an enzyme that cleaves a very large HIV protein into smaller, functional proteins. The virus cannot survive without this enzyme; therefore, it is one of the most effective protein targets for killing HIV.

There are many distributed computing programs, such as the world community grid, which allows people around the world to help scientists by computing calculations. The software adds to the use of super computers by using the unused processing power of millions of home computers. The world community grid works on HIV, cancer, and protein folding. All three projects centre around protein modelling and protein modification models. Using the data gained from distributed computing models of proteins, scientists can develop more specific and effective therapies. In addition, most enzymes act as part of complexes and networks, which also affect the way an enzyme acts in a cell. Understanding these complex networks will assist in developing drugs that affect the function of these complexes.

Biomarkers

Understanding the proteome, the structure and function of each protein and the complexities of protein-protein interactions will be critical for developing the most effective diagnostic techniques and disease treatments in the future.

An interesting use of proteomics is using specific protein biomarkers to diagnose disease. A number of techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked immunosorbent assay (ELISA) or mass spectrometry. The following are some of the diseases that have characteristic biomarkers that physicians can use for diagnosis:

- In Alzheimer's disease, elevations in beta secretase creates amyloid/beta-protein, which causes plaque to build up in the patient's brain, which causes dementia. Targeting this enzyme decreases the amyloid/beta-protein and so slows the progression of the disease. A procedure to test for the increase in amyloid/beta-protein is immunohistochemical staining, in which antibodies bind to specific antigens or biological tissue of amyloid/beta-protein.
- Heart disease is commonly assessed using several key protein based biomarkers. Standard protein biomarkers for CVD include interleukin-6, interleukin-8, serum amyloid A protein, fibrinogen, and troponins. cTnI cardiac troponin I increases in concentration within 3 to 12 hours of initial cardiac injury and can be found elevated days after an acute myocardial infarction. A number of commercial antibody based assays as well as other methods are used in hospitals as primary tests for acute MI.
- Proteomic analysis of kidney cells and cancerous kidney cells is producing promising leads for biomarkers for renal cell carcinoma and developing assays to test for this disease. In kidney-related diseases, urine is a potential source for such biomarkers. Recently, it has been shown that the identification of urinary polypeptides as biomarkers of kidney-related diseases allows to diagnose the severity of the disease several months before the appearance of the pathology.

Branches

1. *Protein separation.* Proteomic technologies rely on the ability to separate a complex mixture so that individual proteins are more easily processed with other techniques.
2. *Protein identification.* Well-known methods include low-throughput sequencing through Edman degradation. Higher-throughput proteomic techniques are based on mass spectrometry, commonly peptide mass fingerprinting on MALDI-TOF instruments, or De novo repeat detection MS/MS on instruments capable of more than one round of mass spectrometry. MS/MS data can be analyzed by simple database searches as is the case for PMFs and additionally they can be analyzed by de novo sequencing and homology searching. This particular approach allows to even identify similar (homolog) proteins, e.g. across species in case a protein was derived from an organism with unsequenced genome. Antibody-based assays can also be used, but are unique to one sequence motif.
3. In quantitative proteomics different methods are used to obtain quantitative information on a proteome-wide scale. Rather than just lists of proteins, quantitative

proteomics provides functional information and reveals temporal changes in the proteome.

4. *Protein sequence analysis* is a branch of bioinformatics that deals with searching databases for possible protein or peptide matches by algorithms such as Mascot, PEAKS(software), OMSSA, SEQUEST and X!Tandem, PWB Protein Identification Cluster Software Solution, functional assignment of domains, prediction of function from sequence, and evolutionary relationships of proteins.
5. *Structural proteomics* concerns the high-throughput determination of protein structures in three-dimensional space. Common methods are x-ray crystallography and NMR spectroscopy.
6. *Interaction proteomics* concerns the investigation of protein interactions on the atomic, molecular and cellular levels. see related article on Protein-protein interaction prediction.
7. *Protein modification* studies the modified forms of proteins. Almost all proteins are modified from their pure translated amino-acid sequence, by so-called post-translational modification. Specialized methods have been developed to study phosphorylation (phosphoproteomics) and glycosylation (glycoproteomics).
8. *Cellular proteomics* is a new branch of proteomics aiming to map the location of proteins and protein-protein interactions in whole cells during key cell events. Centers around the use of techniques such as X-ray Tomography and optical fluorescence microscopy.
9. *Experimental bioinformatics* is a branch of bioinformatics, as it is applied in proteomics, coined by Mathias Mann. It involves the mutual design of experimental and bioinformatics methods to create (extract) new types of information from proteomics experiments.

Technologies

Proteomics uses various technologies:

- One- and two-dimensional gel electrophoresis is used to identify the relative mass of a protein and its isoelectric point.
- X-ray crystallography and nuclear magnetic resonance are used to characterize the three-dimensional structure of peptides and proteins. However, low-resolution techniques such as circular dichroism, Fourier transform infrared spectroscopy and Small angle X-ray scattering (SAXS) can be used to study the secondary structure of proteins.

- Tandem mass spectrometry combined with reverse phase chromatography or 2-D electrophoresis is used to identify proteins using database search tools such as Mascot, Phenyx, PEAKS(software), OMSSA, X!Tandem and SEQUEST or de novo algorithms and quantify all the levels of proteins found in cells.
- Scaffold, a software tool useful in visualization of tandem mass spectrometry results.
- Tandem mass spectrometry combined with tagging technologies such as TMT, ICPL or iTRAQ is used for quantification of proteins and peptides.
- Mass spectrometry (no-tandem), often MALDI-TOF, is used to identify proteins by peptide mass fingerprinting. This technology is also used in so-called "MALDI-TOF MS protein profiling" where samples (i.e. serum) are prepared by either protein chips (SELDI-TOF MS), magnetic beads (The Bruker Daltonics protein profiling platform) or with other methods of sample treatment, such as liquid chromatography, size-exclusion and immunoaffinity. Protein peaks of interest must be identified by tandem mass spectrometry. Protein profiling with MALDI-TOF MS could be of high use in clinical diagnostics, but so far there has been little success with advancing MALDI-TOF MS protein profiling into clinical validation due to high analytical variation.
- ICP-MS combined with MeCAT - Metal Coded Tagging - technology is used for ultrasensitive quantification of proteins and peptides down to low attomol range
- Affinity chromatography, yeast two hybrid techniques, fluorescence resonance energy transfer (FRET), and Surface Plasmon Resonance (SPR) are used to identify protein-protein and protein-DNA binding reactions.
- X-ray Tomography used to determine the location of labeled proteins or protein complexes in an intact cell. Frequently correlated with images of cells from light based microscopes.
- Software based image analysis is utilized to automate the quantification and detection of spots within and among gels samples. While this technology is widely utilized, the intelligence has not been perfected yet. For example, the leading software tools in this area tend to agree on the analysis of well-defined, well-separated protein spots, but they deliver different results and tendencies with less-defined less-separated spots - thus necessitating manual verification of results.

Branches of proteomics

Structural proteomics

Structural proteomics includes the analysis of protein structures at large-scale. It compares protein structures and helps identify functions of newly discovered genes.

The structural analysis also helps to understand that where drugs bind to proteins and also show where proteins interact with each other. This understanding is achieved using different technologies such as X-ray crystallography and NMR spectroscopy.[37]

Expression proteomics

Expression proteomics includes the analysis of protein expression at larger scale. It helps identify main proteins in a particular sample, and those proteins differentially expressed in related samples—such as diseased vs. healthy tissue. If a protein is found only in a diseased sample then it can be a useful drug target or diagnostic marker. Proteins with same or similar expression profiles may also be functionally related. There are technologies such as 2D-PAGE and mass spectrometry that are used in expression proteomics.[37]

Interaction proteomics

Interaction proteomics is the analysis of protein interactions at larger scale. The characterization of protein-protein interactions are useful to determine the protein functions and it also explains the way proteins assemble in bigger complexes. Technologies such as affinity purification, mass spectrometry, and the yeast two-hybrid system are particularly useful in interaction proteomics.

Protein-Protein Interactions

Definition

Specific interactions between two or more proteins.

Examples

Enzyme-inhibitor complex; antibody-antigen complex; receptor-ligand interactions, multiprotein complexes such as ribosomes or RNA polymerases

Characteristics

Classification: Protein-protein interactions can be arbitrarily classified based on the proteins involved (structural or functional groups) or based on their physical properties (weak and transient, “non-obligate” vs. strong and permanent). Protein interactions are usually mediated by defined domains, hence interactions can also be classified based on the underlying domains.

Universality: All of molecular biology is about protein-protein interactions (Alberts et al. 2002, Lodish et al. 2000). Protein-protein interactions affect all processes in a cell: structural proteins need to interact in order to shape organelles and the whole cell, molecular machines such as ribosomes or RNA polymerases are held together by protein-protein interactions, and the same is true for multi-subunit channels or receptors in membranes.

Specificity distinguishes such interactions from random collisions that happen by Brownian motion in the aqueous solutions inside and outside of cells. Note that many proteins are known to interact although it remains unclear whether certain interactions have any physiological relevance.

Number of interactions: It is estimated that even simple single-celled organisms such as yeast have their roughly 6000 proteins interact by at least 3 interactions per protein, i.e. a total of 20,000 interactions or more. By extrapolation, there may be on the order of ~100,000 interactions in the human body.

Protein-protein interactions and protein complexes: Most protein-protein interactions are detected as interacting pairs or as components of protein complexes. Such complexes may

contain dozens or even hundreds of protein subunits (ribosomes, spliceosomes etc.). It has even been proposed that all proteins in a given cell are connected in a huge network in which certain protein interactions are forming and dissociating constantly

Structural features of protein-interaction sites

Hundreds of protein complexes have been analyzed by X-Ray crystallography and other methods. Data about the structures of proteins and complexes are available from the Protein Databank (PDB, <http://www.rcsb.org/>). The following statements about the geometry and energetics of protein interactions have been drawn from the analysis of several dozens to about a 100 protein pairs and complexes that have been crystallized.

The contact area between two proteins is almost always bigger than 1100 Å² with each of the interacting partners contributing at least 550 Å² of complementary surface. On average each partner loses about 800 Å² of solvent-accessible surface upon contact, contributed by some 20 amino acid residues of each partner, i.e. the average interface residue covers some 40 Å². On average, dimers contribute 12% of their accessible surface area to the contact interface, trimers 17.4% and tetramers 20.9%. However, variations are large and total interface areas range from 6% in inorganic pyrophosphatase homodimers to 29% in Trp repressor homodimers. For data sets of 10 enzyme-inhibitor complexes and 6 antibody-antigen complexes the mean interface area was about 780 Å².

Forces that mediate protein-protein interactions include electrostatic interactions, hydrogen bonds, the van der Waals attraction and hydrophobic effects. The average protein-protein interface is not less polar or more hydrophobic than the surface remaining in contact with the solvent. Water is usually excluded from the contact region. Non-obligate complexes tend to be more hydrophilic in comparison, as each component has to exist independently in the cell. It has been proposed that hydrophobic forces drive protein-protein interactions and hydrogen bonds and salt bridges confer specificity. *Van der Waals interactions* occur between all neighbouring atoms, but these interactions at the interface are no more energetically favourable than those made with the solvent. However, they are more numerous, as the tightly packed interfaces are more dense than the solvent and hence they contribute to the binding energy of association. *Hydrogen bonds* between protein molecules are more

favourable than those made with water. Interfaces in permanent associations tend to have fewer hydrogen bonds

than interfaces in non-obligate associations. The number of hydrogen bonds is about 1 per 170 Å² buried surface. A standard size interface (~ 1600 Å²) buries about 900 Å² of the non-polar surface, 700 Å² of polar surface, and contains 10 (\pm 5) hydrogen bonds. In a set of reasonably stable dimers there are, on average, 0.9 to 1.4 hydrogen bonds per 100 Å² of contact area buried (interfaces usually covering > 1000 Å²), but

the number of hydrogen bonds varies from 0 in some complexes (e.g. uteroglobin) to 46 in variant surface glycoprotein. Side-chain hydrogen bonds represent approximately 76-78% of the interactions. Only 56% of homodimers were found to possess *salt bridges* (many having none, and at the most five).

Shape: Independent studies showed that 83-84% of interfaces are more or less flat. With few exceptions, the interfaces are approximately circular areas on the protein surface in both permanent and non-obligate complexes. Interfaces in permanent associations tend to be larger, less planar, more highly segmented (in terms of sequence), and closer packed than interfaces in non-obligate associations.

Complementarity: can be measured in terms of “fitting surface shape”. Interfaces in homodimers, enzyme-inhibitor complexes, and permanent heterocomplexes are the most complementary, whilst the antibody-antigen complexes and the non-obligate heterocomplexes are the least complementary.

Secondary structure: In one study the loop interactions contributed, on average, 40% of the interface contacts. In another study (involving 28 homodimers), 53% of the interface residues were α -helical, 22% β sheets, and 12% ab, with the rest being coils.

Amino acid composition: Interfaces have been shown to be more hydrophobic than the exterior but less hydrophobic than the interior of a protein. In one study, 47% of interface residues were hydrophobic, 31% polar and 22% charged. Permanent complexes have interfaces that contain hydrophobic residues, whilst the interfaces in 5 non-obligate complexes favour the more polar residues. Site-directed mutagenesis showed that in many cases a large majority (i.e. > 50%) of interface residues can be mutated to alanine with little effect on K_d: i.e. the functional epitope is a subset of the structural epitope.

Thermodynamics

Protein-protein interactions can be described as simple chemical reactions of the form where A and B are two proteins and AB is a protein complex, i.e. an interacting pair of proteins. Multiprotein complexes are usually thought to be assembled by adding subunits successively.

Qualitative description of stability: Protein interactions can be weak and extremely short-lived (“non-obligate”) or strong and permanent (i.e. “stable”). For example, an enzyme might bind a protein substrate in order to phosphorylate it and then dissociate after less than a microsecond. Other protein complexes such as the triple-helix of collagen may reside in bones and other tissues for weeks or even years without dissociation.

Quantitative description of protein-protein interactions

Quantitatively, the interaction of two proteins follows the mass action law:

k_a = second-order rate constant for the bimolecular association reaction,

k_d = first-order rate constant for the uni-molecular dissociation reaction,

$K_d = k_d / k_a$ = equilibrium constant for dissociation (K_a for association). K_d is related to the concentrations of A, B, and AB at thermodynamic equilibrium. K_d has the dimension of a concentration and is expressed in mole/litre (or “M”). The range of values observed in biologically relevant processes that rely on protein-protein

interactions is extremely wide, and extends over at least 12 orders of magnitude from 10^{-4} to 10^{-16} M.

K_d values in the mM range are considered as rather weak, values in the nM range or below as strong (e.g. trypsin – pancreatic trypsin inhibitor [PTI] has a dissociation constant in the order of 10^{-14} M). However, the biological strength may depend on other effects such as cooperativity. For example, several weak interactions between the subunits of a complex may still result in a highly stable complex.

Energetics: With a K_d ranging from 10^{-4} to 10^{-14} M, the free enthalpy of dissociation ΔG_d ranges from 6 to 19 kcal/mol, i.e. 19 kcal are required to separate 1 mole of trypsin-PTI. The free enthalpy can be broken down into an enthalpic and an entropic contribution. In any case, dehydration of the non-polar groups at the interface is essential for stable association. Actual K_d values can be retrieved from some of the databases mentioned below such as the Database of Interacting Proteins, DIP. A single pairwise

interaction between amino acids may account for as much as 6 kcal/mol. Residue pairs that form salt bridges and charged hydrogen bonds yield the largest values; pairs making neutral hydrogen bonds or non-polar interactions are in the 0-3 kcal/mol range. This is much less than the energy of a hydrogen bond, and implies that interaction between the two residues in the complex is only marginally stronger than the interactions with water they make in the free proteins. In complexes of known 3D structure, the peptide group is part of at least half the hydrogen bonds at protein-protein interfaces. Side chain to main chain bonds are especially common, and main chain to main chain bonds do occur. It has been estimated from an empirical correlation that for non-polar surfaces (i.e. hydrophobic interactions), there is an energy gain of approximately 25 to 50 calories per Å² (up to 72 calories per Å² based on other studies). Many dimeric interactions, that can have association constants greater than 10¹⁶ M⁻¹, are so strong that the monomers have to be denatured to separate the complex.

Methods to study protein-protein interactions

Details on the methods for the analysis of protein-protein interactions can be found in other sections of this Encyclopedia. Important experimental techniques are listed

Purification of protein complexes and their separation by 2-dimensional gel electrophoresis or liquid chromatography. After separation the components of a protein complex can be identified by the staining of protein bands by various dyes such as Coomassie Blue. Furthermore, the protein bands can be analyzed by mass spectrometry which can also be used to sequence proteins.

In-vitro binding experiments involve purified proteins whose interaction is detected by retention of one partner by a second protein that has been immobilized on some sort of matrix (such as agarose or glass).

In vivo methods include various **two-hybrid systems**, **FRET**, and other systems that usually require that proteins to be studied are bound to some other detectable protein or chemical compound.

Dissociation constants in the micromolar range can be determined by **equilibrium ultracentrifugation**, **microcalorimetry** and other methods. When K_d is in the nanomolar range, radioisotope labelling or ELISA techniques can be used.

Site directed mutagenesis is used to identify the role of individual amino acids.

Structural methods such as **NMR** or **X-ray crystallography**

Yeast Two hybrid analysis

Purpose

The yeast "two hybrid" or "interaction trap" assay, is a technique used to study protein-protein interactions, which are critical in virtually all cellular processes. The study of protein-protein interactions can be divided into three major parts:

- identification of binding proteins
- the characterization of known interactions
- the potential to manipulate such interactions¹

Theory

The yeast two hybrid assay, developed by Fields and Colleagues over a decade ago², has since become one of the most popular tools used in molecular biology. In order to understand how the two hybrid system was first developed, it is useful to understand the process of galactose metabolism in yeast.

Galactose metabolism

Galactose is imported into the cell and converted to galactose-6-phosphate by six enzymes (GAL1, GAL2, PGM2, GAL7, GAL10, MEL1) which are transcriptionally regulated by the proteins Gal80, Gal3, and Gal4, the latter of which plays the central role of DNA-binding transactivator (Figure 1). Gal80 binds Gal4 and inhibits its transcriptional ability. Gal3, in the presence of galactose, binds and causes a conformational change in Gal80, which then allows Gal4 to function as a transcriptional activator

Gal4, like other transcriptional activators, is a modular protein that requires both DNA-binding (BD) and activation domains (AD). The "two hybrid" technique exploits the fact that Gal4 cannot function as a transcriptional activator unless physically bound to an activation domain. Furthermore, it has been demonstrated that this interaction does not need to

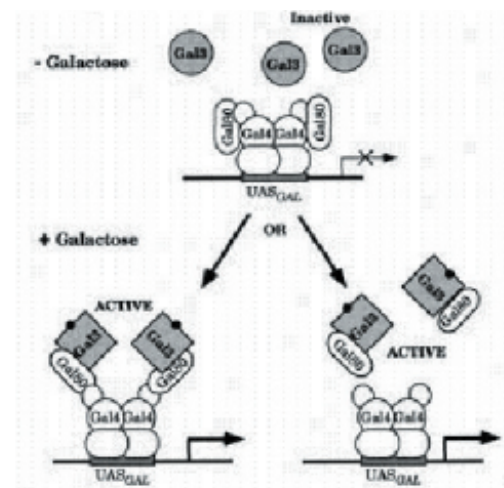


Figure 1 Galactose metabolism

be covalent: an experiment was performed where the negative regulatory protein, Gal80, was fused with an activation domain to produce Gal80-AD, and was able to act as a transcriptional activator through its natural binding interaction with a Gal4 protein that was missing its own activation domain.

Based on the above observations, a two hybrid assay is performed by expressing two fusion proteins in yeast, the "hunter" and the "bait", where the "hunter" protein is the potential binding partner fused to a yeast activation domain, and the bait is your protein of interest is fused to a yeast binding domain.

The hunter and bait constructs are co-transfected into the yeast strain, containing the appropriate "Upstream Activating Sequence" (UAS) proximal to a reporter gene, which is expressed if a binding interaction occurs between the hunter and the bait (Figure 3).

To identify proteins that may interact with the bait, a plasmid library, expressing cDNA-encoded AD-fusion proteins, can be screened by introduction into a yeast

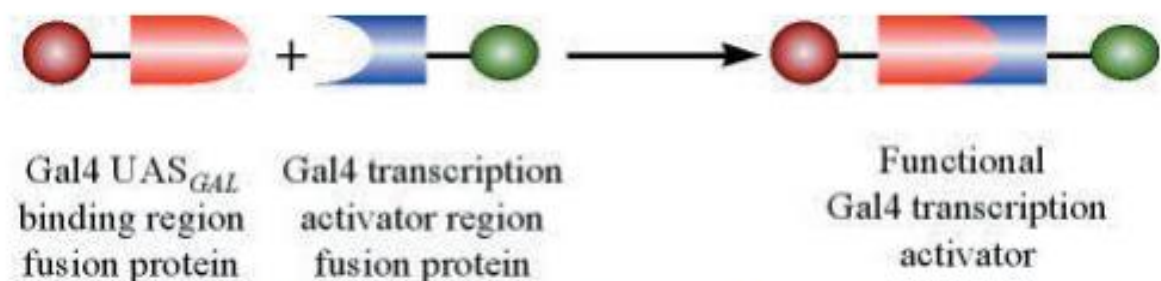


Figure 2 Gal4 transcriptional activator exploited for yeast two hybrid assay

strain. The interaction of the potential hunter proteins with the bait results in the activation of the reporter gene, allowing for the identification of cells containing the interacting proteins.

The Pros and Cons of the Two hybrid Assay

Disadvantages and Trouble-Shooting:

- 1) Since the two hybrid assay measures reporter activity in response to transcriptional activation, an obvious problem would arise if your protein of interest were able to activate transcription on its own (auto-activation). It is, therefore, imperative that an initial experiment be done to test for the transcriptional activity of the protein of interest.
- 2) The extensive use of fusion proteins may change the conformation of the hunter or bait, which may alter activity or binding. Nevertheless, there have been shown to be few problems with tagged proteins, perhaps due to their modular nature, where domains can fold independently, often allowing the introduction of artificial modules.
One way to test whether the protein of interest is folded properly is to clone a positive interactor (a protein known to interact with the bait) into the vector and test for a two hybrid interaction, which will only result if both proteins are folded correctly. However, this technique only works for the domains involved in "positive" interactions and may not be conclusive for domains involved in novel interactions. Furthermore, an empirical way to circumvent this problem may be through the reciprocal transfer of proteins, involving the switching of a BD-fusion to an AD-fusion protein and vice et versa.
- 3) A major drawback of testing protein-protein interactions in a heterologous system such as the yeast is that interactions may depend on certain post-transcriptional modifications, such as disulfide bridge formation, glycosylation, or phosphorylation, which may not occur properly or at all in the yeast system¹.
This problem may be, in certain cases, circumvented perhaps by co-expressing the enzymes

necessary for such modifications.

- 4) Since the fusion proteins in the two hybrid system must be targeted to the nucleus, extracellular proteins or proteins with stronger targeting signals may be at a disadvantage.
- 5) In the classical two hybrid library preparations, only one sixth of the cDNA is in the correct frame. Nevertheless, this works out to over a million independent clones to be studied (which pushes feasibility), if a good representation is to be obtained.
A solution to this may be to make directional libraries of a relevant tissue or cell type, or if possible, to simply choose a less complex organism to study.
- 6) It has been shown that sub-domains of proteins may interact better than full length clones, perhaps due to the lack of certain folding restraints. Since screening of libraries selects for optimized reactions, one may obtain a false representation. This problem can be dealt with by very tediously ensuring that only full-length cDNAs in the correct reading frame are cloned.
- 7) Given that the two hybrid assay measures reporter activity, it cannot be excluded that a third protein may perhaps be bridging the hunter and the bait. This possibility is unlikely, but should always be kept in mind, since this is also a problem encountered with many other conventional biochemical binding assays.
- 8) Certain proteins, when expressed in the yeast system or targeted to the nucleus, may become toxic. Other proteins may degrade essential yeast proteins or proteins whose presence are required for the assay. Such genes may be counter-selected for during growth and may result in problems.
- 9) As typical with all exhaustive screening assays, the identification of false binding partners presents itself as a disadvantage in the two hybrid assay. Due to localization and time restraints, and cell context in different cell types, even though two proteins can interact, it is not certain that they will interact in real situations given the factors above. Therefore, the biological relevance of any two proteins found to interact is a great concern,

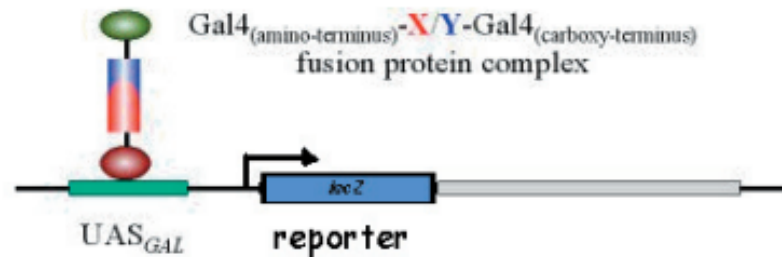


Figure 3 Gal4 binds to an upstream activating sequence

and must be kept in perspective.

Advantages:

- 1) The yeast two hybrid system has a clear advantage over classical biochemical or genetic methods, in that it is an *in vivo* technique that uses the yeast cell as a living test-tube.
- 2) The use of the yeast host can be considered an advantage since it bears a greater resemblance to higher eukaryotic systems than a system based on a bacterial host.
- 3) With regards to classical biochemical approaches, which can require high quantities of purified proteins or good quality anti-bodies, the two hybrid system has minimal requirements to initiate screening, since only the cDNA of the gene of interest is needed.
- 4) In signaling cascades, weak and transient interactions are often very important. Such interactions are more readily detected with the two hybrid system since the reporter gene response often leads to significant amplification. However, one must keep in mind that there are trade-offs between detecting weak signal and obtaining false-positives in screening procedures.
- 5) The two hybrid assay is also useful for analysis of known interactions, which can be achieved by modifying important residues or modules and observing this effect on binding.
- 6) Interactions can be measured semi-quantitatively using the two hybrid system, allowing discrimination between high, intermediate, and low-affinity bindings, the power of which correlates with that of *in vitro* approaches.
- 7) Although the two hybrid assay was predicted to be limited to the study of cellular proteins, given that extracellular proteins often undergo modifications such as glycosylation or disulfide cross-links that are not expected to occur in the yeast nucleus, there have been various reported successes with extracellular receptor/ligand complexes.
- 8) Two hybrid screens are sometimes termed "functional screens", since if at least one of the proteins screened has a known function in a well-defined

pathway, it might provide a functional hint in the current interaction.

- 9) Although there are certain disadvantages involving the two hybrid assay, the most convincing argument for its use is the speed and ease by which the molecular mechanisms of many signaling cascades have been defined using this technique.

Methods and Protocols

There are numerous two hybrid systems that can be used to detect protein-protein interactions for different purposes. Here are listed some of the more common methods used for such studies, as well as a brief outline of practical steps to help maximize experimental success. Some protocols for these methods are also provided in the links provided.

Standard Yeast Two hybrid Systems

- 1) The Gal4 System identifies the interaction between two proteins by reconstituting active Gal4 protein. The two proteins involved are expressed as fusion proteins with the Gal4 BD and AD. The two plasmids containing these constructs are co-transfected into a strain of yeast containing the upstream activation sequences from the GAL1-GAL10 regions, which promote transcription of the *E. coli* LacZ gene. If interaction occurs, LacZ is transcribed, resulting in the turning blue of the strain when placed in medium containing X-gal (a chromogenic substrate). Note that when using the Gal4 system, to avoid interference by endogenous Gal4 and Gal80 proteins, the yeast host must carry deletions in the GAL4 GAL8 genes, and due to the deletion of these genes, the yeast cells will grow more slowly than wild type⁷.
- 2) The LexA system is another version of the two hybrid system that uses the operator sequence and the BD from the *E. coli* LexA repressor protein. In this system, the AD comes from a segment of *E. coli* DNA that expresses an acidic peptide, acting as transcriptional activator in yeast when fused with a DNA binding domain. These constructs are then co-transfected into yeast strains containing the LexA operators upstream of either the *E. coli*

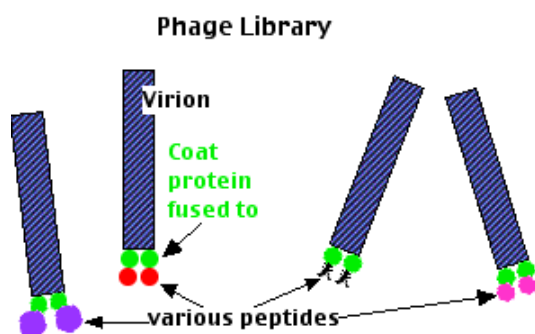
Phage Display

This method exploits:

- a DNA bacteriophage that infects *E. coli*;
- its ability to remain infectious even if one of its coat proteins contains segments of a foreign protein.

The method:

1. Transform bacteriophages with a
 - random mix of DNA from the organism you are interested in coupled to
 - the DNA encoding one of the viral coat proteins.
2. Infect *E. coli* with these phages.
3. As the viruses replicate, they will not only propagate the recombinant gene but also express it as a coat protein.
4. Both will be incorporated into new virions.
5. Harvest the mix of viruses.
6. Pass the mixture through an affinity chromatography column to which your "target"



protein has been fixed.

7. Those viruses that display a piece of foreign protein (peptide) that can bind to the target will stick to it.
8. Elute the bound phage with a buffer.
9. Repeat steps 6–8 to further enrich your binders.
10. Infect *E. coli*.
11. Grow separate colonies (clones).
12. Sequence the coat protein gene to find the sequence of the foreign DNA inserted in it.
13. Using the codon table, determine the amino acid sequence of the peptide.
14. Search databases for a protein containing this sequence.
15. Result: another protein that associates with your target protein.

Phage display is also used to make monoclonal antibodies (without the need for mice).

Regulation of protein-protein interactions

Interactions between proteins are tightly regulated. The most important ways of regulation involve the following mechanisms:

Expression: Only proteins that are expressed in the same place can interact. RNA and protein expression (i.e. transcription and translation) are therefore regulating such interactions in a time- and space-dependent manner. For example, various Fibroblast Growth Factors (FGFs) are expressed in different tissues such as limbs, brain, kidney etc. while their cognate receptors are similarly expressed in particular tissues where their interaction activates signalling pathways that in turn lead to the expression of certain target genes. FGFs are also expressed in a time-dependent manner, e.g. FGF4 and 8 are only expressed in embryos while others are mostly expressed in adult tissues. The same is true for FGF receptors.

Modifications: Covalent modification of proteins can prevent or induce binding of others. Examples are proteins that contain the so-called bromo domain that binds only acetylated but not unmodified histones.

Ligands: Small molecules or other proteins can regulate protein interactions. For example, binding of GTP to trimeric G proteins promotes their dissociation while binding of GDP favors their association.

Subcellular Localization: Several transcription factors such as NFkB are present in the cytoplasm or even in membranes (such as Notch). They can be activated by translocation to the nucleus where they interact with other transcription factors in order to activate gene transcription. Transport is also important for proteins that are exported from cells such as peptide hormones.

Stability: As for expression, proteins can only interact when they are present. Regulated degradation of proteins prevents them from further interactions.

Prediction of protein-protein interactions

Prediction of interaction sites in proteins of known structure usually focuses on the location of hydrophobic surface clusters on proteins. In one study, this method predicted the correct interaction site in 25 out of 29 cases. Other methods include solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area. Among a test set of 28 homodimers, the known interface site was found to be amongst the most planar, protruding, and accessible patches, and amongst the patches with highest interface propensity. Nevertheless, one of the algorithms (PATCH) that

uses multiple parameters predicted the location of interface sites in known complexes only for 66% of the structures.

Prediction of interacting protein pairs from genome sequences

Several attempts have been made to predict protein-protein interactions from genome sequences. The major methods are:

Rosetta stone proteins: Some protein sequences have been found to be split into two independent proteins in other organisms. From such sequences it has been concluded that the two independent proteins form a complex, based on the (covalent) association in the former organism. Such fusion proteins are called Rosetta stone proteins. Supposedly they predict interactions among related proteins. Example: human succinyl CoA transferase is split in *E. coli* into acetate CoA transferases a and b subunits.

Phylogenetic profiles: Some protein pairs are evolutionarily maintained together in many different organisms. It has been concluded that such “co-evolving” proteins are associated either functionally or even physically, i.e. by a protein-protein contact.

Clinical relevance and applications of protein-protein interaction analysis

Biologically active proteins such as peptide hormones or antibodies act by interacting with other proteins such as receptors or antigens, respectively. Knowing their interaction sites allows the modification of the activity of such proteins or changing their specificity. In addition, small molecules may be designed that block interactions such as the binding of virus coat proteins to their cellular receptors, thereby blocking infection. Proteins and their interactions are therefore potential **drug targets**. Sometimes, protein-protein interactions are disadvantageous, such as in insulin that tends to form dimers and hexamers which are less active than monomers. Genetically engineered insulin molecules retain biological activity without oligomerizing.

Importance of Protein Interactions

The study of protein interactions has been vital to the understanding of how proteins function within the cell. Publication of the draft sequence of the human genome and proteomics-based protein profiling studies catalyzed a resurgence in protein interaction analysis. Characterizing the interactions of proteins in a given cellular proteome (now often referred to as the “interactome”) will be the next milestone along the road to understanding the biochemistry of the cell.

The ~30,000 genes of the human genome are speculated to give rise to 1×10^6 proteins through a series of post-translational modifications and gene splicing mechanisms. Although a population of these proteins can be expected to work in relative isolation, the majority are expected to operate in concert with other proteins in complexes and networks to orchestrate the myriad of processes that impact cellular structure and function. These processes include cell cycle control, differentiation, protein folding, signaling, transcription, translation, post-translational modification and transport.

Implications about function can be made via protein:protein interaction studies. These implications are based on the premise that the function of unknown proteins may be discovered if captured through their interaction with a known protein target of known function.

Consequences of Protein Interactions

The result of two or more proteins interacting with a specific functional objective can be demonstrated in several different ways. The measurable effects of protein interactions have been outlined by Phizicky and Fields. Protein interactions can:

- Alter the kinetic properties of enzymes. This may be the result of subtle changes at the level of substrate binding or at the level of an allosteric effect.
- Allow for substrate channeling by moving a substrate between or among subunits, resulting ultimately in an intended end product.
- Create a new binding site, typically for small effector molecules.
- Inactivate or destroy a protein.
- Change the specificity of a protein for its substrate through interaction with different binding partners; e.g., demonstrate a new function that neither protein can exhibit alone.
- Serve a regulatory role in either an upstream or a downstream action.

Types of Protein Interactions

Protein interactions fundamentally can be characterized as stable or transient. Both stable and transient interactions can be either strong or weak. Stable interactions are those associated with proteins that are purified as multi-subunit complexes. The subunits of the complex can be identical or different. Hemoglobin and core RNA polymerase are two examples of stable multi-subunit complex interactions. Stable interactions are best studied by co-immunoprecipitation, pull-down or far-Western methods.

Transient interactions are expected to control the majority of cellular processes. As the name implies, transient interactions are on/off or temporary in nature and typically require a set of conditions that promote the interaction. Transient interactions can be strong or weak, fast or slow. While in contact with their binding partners, transiently interacting proteins are expected to be involved in the whole range of cellular processes including protein modification, transport, folding, signaling, cell cycling, etc. Transient interactions can be captured by cross-linking or label transfer methods.

Factors Affecting Protein Protein Interactions

The analysis of protein protein interactions can be of two kinds :

1. Qualitative
2. Quantitative

In both the cases we have factors related to the interactions. Factors affecting the qualitative analysis are discussed before. In this section we would mainly try to concentrate on the quantitative analysis.

Factors Affecting the Quantitative protein protein interactions

With regards to binding factors we mean whether the interaction is one-to-one or one-to-many binding. The one-to-one binding interactions can be quantified very easily but the problems lie when one protein has more than one binding sites. For one-to-many binding sites, there can be again two possibilities. If all these many binding are not associated with each other or in other words independent of each other we can just treat each binding as one-to-one and then sum them up all together. But if all these interactions are dependent on each other then we have to deal with its allosteric effects.

So the factors affecting the binding properties are :

1. Affinity
2. Kinetics
3. Thermodynamics Properties

Affinity

The dissociation constant is commonly used to describe the affinity between a ligand (L) (such as a drug) and a protein (P) i.e. how tightly a ligand binds to a particular protein. The dissociation constant is usually denoted K_d and is the inverse of the association constant. Ligand-protein affinities are influenced by non-covalent intermolecular interactions between the two molecules such as hydrogen bonding, electrostatic interactions, hydrophobic and Van der Waals forces. They can also be affected by high concentrations of other macromolecules, which causes macromolecular crowding.

1. K_a (affinity constant) or K_d (dissociation Constant)
2. $K_d = 1/K_a$

Kinetics

kinetics is a branch of chemical kinetics in which the kinetic species are defined by different non-covalent bindings and/or conformations of the molecules involved, which are denoted as *receptor(s)* and *ligand(s)*.

A main goal of receptor-ligand kinetics is to determine the concentrations of the various kinetic species (i.e., the states of the receptor and ligand) at all times, from a given set of initial concentrations and a given set of rate constants. In a few cases, an analytical solution of the rate equations may be determined, but this is relatively rare. However, most rate equations can be integrated numerically, or approximately, using the steady-state approximation. A less ambitious goal is to determine the final *equilibrium* concentrations of the kinetic species, which is adequate for the interpretation of equilibrium binding data.

A converse goal of receptor-ligand kinetics is to estimate the rate constants and/or dissociation constants of the receptors and ligands from experimental kinetic or equilibrium data. The total concentrations of receptor and ligands are sometimes varied systematically to estimate these constants.

1. $k(\text{ass})$ or $k(\text{on})$ association rate constant or on rate
2. $K(\text{diss})$ or $K(\text{off})$ dissociation rate constant or off rate

Thermodynamics properties

1. Enthalpy Change
2. Entropy Change
3. Heat Capacity change
4. Gibbs free energy change
5. Role of Water

Role of Water

Since most of the protein protein interactions take place in aqueous phase we should also consider the affect of water in an interaction. Water makes the environment quite hydrophobic and considerably favourable for protein interactions as they are also hydrophobic molecules. Due to the abundance of water molecules, most of the bonds need to be broken to make the protein interactions favorable. Water can also act as a cementing material in the hydrophobic gaps between two proteins.

Protein DNA interactions

Protein–DNA interactions are when a protein binds a molecule of DNA, often to regulate the biological function of DNA, usually the expression of a gene. Among the proteins that bind to DNA are transcription factors that activate or repress gene expression by binding to DNA motifs and histones that form part of the structure of DNA and bind to it less specifically. Also proteins that repair DNA such as uracil-DNA glycosylase interact closely with it. In general, proteins bind to DNA in the major groove; however, there are exceptions.^[1] Protein–DNA interaction are of mainly two types, either specific interaction, or non-specific interaction.

Detection methods

There are many in vitro and in vivo techniques which are useful in detecting DNA-Protein Interactions. The following lists some methods currently in use:[3]

- Electrophoretic mobility shift assay is a widespread technique to identify protein-DNA interactions.
- DNase footprinting assay can be used to identify the specific site of binding of a protein to DNA.
- Chromatin immunoprecipitation is used to identify the sequence of the DNA fragments which bind to a known transcription factor. This technique when combined with high throughput sequencing is known as ChIP-Seq and when combined with microarrays it is known as ChIP-chip.
- Yeast One-hybrid System (Y1H) is used to identify which protein binds to a particular DNA fragment.
- Bacterial one-hybrid system (B1H) is used to identify which protein binds to a particular DNA fragment.
- Structure determination using X-ray crystallography has been used to give a highly detailed atomic view of protein-DNA interactions

The Importance of DNA-binding Proteins

As in all areas of molecular biology and genetics, the amount we know about a topic depends on the range and effectiveness of the methods available for its study. With regard to DNA-binding proteins we are fortunate in having a number of powerful techniques that can provide information on the interaction between a protein and the DNA sequence or sequences that it binds to. These techniques can be divided into three categories:

- Methods for identifying the region(s) of a DNA molecule to which a protein binds;
- Methods for purifying a DNA-binding protein;
- Methods for studying the tertiary structure of a DNA-binding protein, including the complex formed when the protein is bound to DNA.

9.1.1. Locating the positions of DNA-binding sites in a genome

Often the first thing that is discovered about a DNA-binding protein is not the identity of the protein itself but the features of the DNA sequence that the protein recognizes. This is because genetic and molecular biology experiments, which we will deal with later in this chapter, have shown that many of the proteins that are involved in genome expression bind to short DNA sequences immediately upstream of the genes on which they act (Figure 9.2). This means that the sequence of a newly discovered gene, assuming that it includes both the coding DNA and the regions upstream of it, provides immediate access to the binding sites of at least some of the proteins responsible for expression of that gene. Because of this, a number of methods have been developed for locating protein binding sites within DNA fragments up to several kb in length, these methods working perfectly well even if the relevant DNA-binding proteins have not been identified.

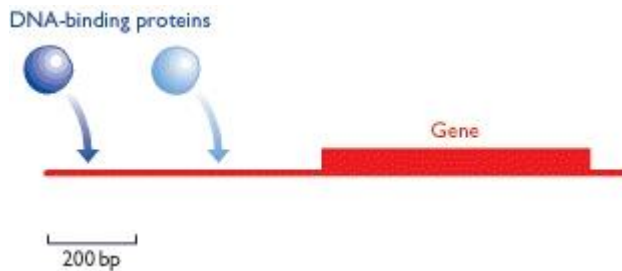


Figure 9.2

Attachment sites for DNA-binding proteins are located immediately upstream of a gene. See Sections 9.2 and 9.3 for more information on the location and function of these protein attachment sites.

Gel retardation identifies DNA fragments that bind to proteins

The first of these methods makes use of the substantial difference between the electrophoretic properties of a ‘naked’ DNA fragment and one that carries a bound protein. Recall that DNA fragments are separated by agarose gel electrophoresis because smaller fragments migrate through the pore-like structure of the gel more quickly than do larger fragments (see Technical Note 2.1). If a DNA fragment has a protein bound to it then its mobility through the gel will be impeded: the DNA-protein complex therefore forms a band at a position nearer to the starting point (Figure 9.3). This is called **gel retardation** (Garner and

Revzin, 1981). In practice the technique is carried out with a collection of restriction fragments that span the region thought to contain a protein binding site. The digest is mixed with an extract of nuclear proteins (assuming that a eukaryote is being studied) and retarded fragments are identified by comparing the banding pattern obtained after electrophoresis with the pattern for restricted fragments that have not been mixed with proteins. A nuclear extract is used because at this stage of the project the DNA-binding protein has not usually been purified. If, however, the protein is available then the experiment can be carried out just as easily with the pure protein as with a mixed extract.

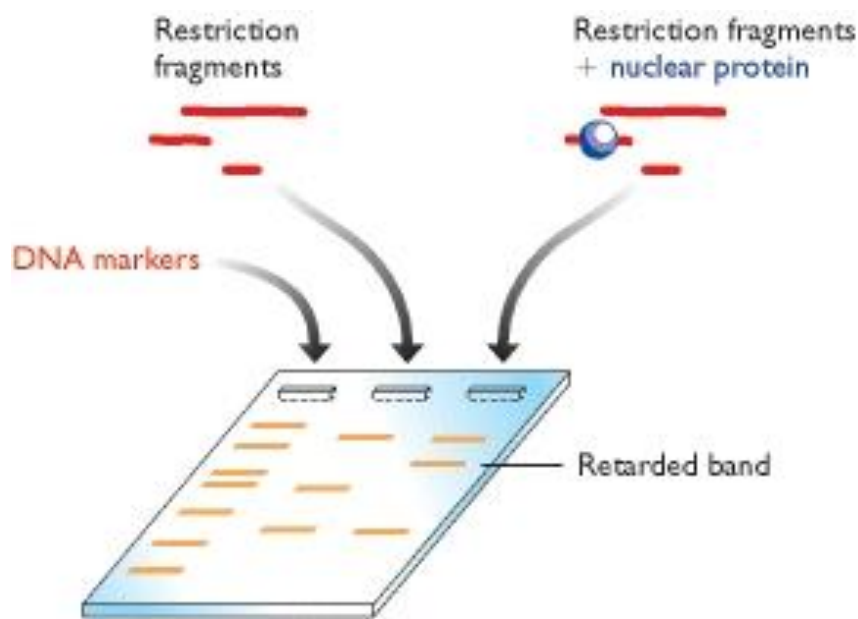


Figure 9.3

Gel retardation analysis. A nuclear extract has been mixed with a DNA restriction digest and a DNA-binding protein in the extract has attached to one of the restriction fragments. The DNA-protein complex has a larger molecular mass than the 'naked' (more...)

Protection assays pinpoint binding sites with greater accuracy

Gel retardation gives a general indication of the location of a protein binding site in a DNA sequence, but does not pinpoint the site with great accuracy. Often the retarded fragment is several hundred bp in length, compared with the expected length of the binding site of a few tens of bp at most, and there is no indication of where in the retarded fragment the binding site lies. Also, if the retarded fragment is long then it might contain separate binding sites for several proteins, or if it is quite small then there is the possibility that the binding site also includes nucleotides on adjacent fragments, ones that on their own do not form a stable complex with the protein and so do not lead to gel retardation. Retardation

studies are therefore a starting point but other techniques are needed to provide more accurate information.

Modification protection assays can take over where gel retardation leaves off. The basis of these techniques is that if a DNA molecule carries a bound protein then part of its nucleotide sequence will be protected from modification. There are two ways of carrying out the modification:

- By treatment with a nuclease, which cleaves all phosphodiester bonds except those protected by the bound protein;
- By exposure to a methylating agent, such as dimethyl sulfate which adds methyl groups to G nucleotides. Any Gs protected by the bound protein will not be methylated.

The practical details of these two techniques are shown in Figures 9.4 and 9.5. Both utilize an experimental approach called footprinting. In nuclease footprinting (Galas and Schmitz, 1978), the DNA fragment being examined is labeled at one end, complexed with binding protein (as a nuclear extract or as pure protein), and treated with deoxyribonuclease I (DNase I). Normally, DNase I cleaves every phosphodiester bond, leaving only the DNA segment protected by the binding protein. This is not very useful because it can be difficult to sequence such a small fragment. It is quicker to use the more subtle approach shown in Figure 9.4. The nuclease treatment is carried out under limiting conditions, such as a low temperature and/or very little enzyme, so that on average each copy of the DNA fragment suffers a single 'hit' - meaning that it is cleaved at just one position along its length. Although each fragment is cut just once, in the entire population of fragments all bonds are cleaved except those protected by the bound protein. The protein is now removed, the mixture electrophoresed, and the labeled fragments visualized. Each of these fragments has the label at one end and a cleavage site at the other. The result is a ladder of bands corresponding to fragments that differ in length by one nucleotide, the ladder broken by a blank area in which no labeled bands occur. This blank area, or 'footprint', corresponds to the positions of the protected phosphodiester bonds, and hence of the bound protein, in the starting DNA.

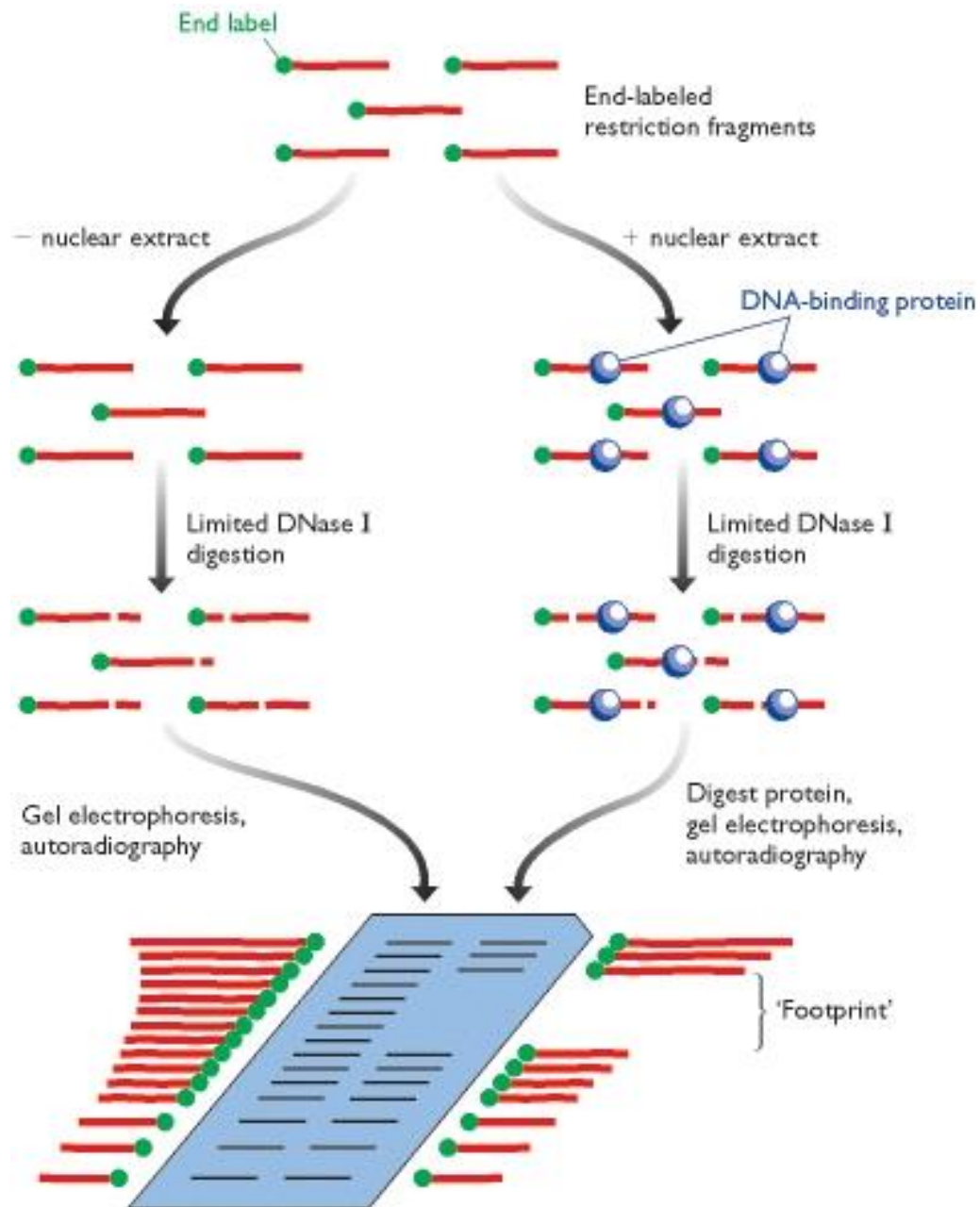


Figure 9.4

DNase I footprinting. The technique is described in the text. The restriction fragments used at the start of the procedure must be labeled at just one end. This is usually achieved by treating a set of longer restriction fragments with an enzyme that(more...)

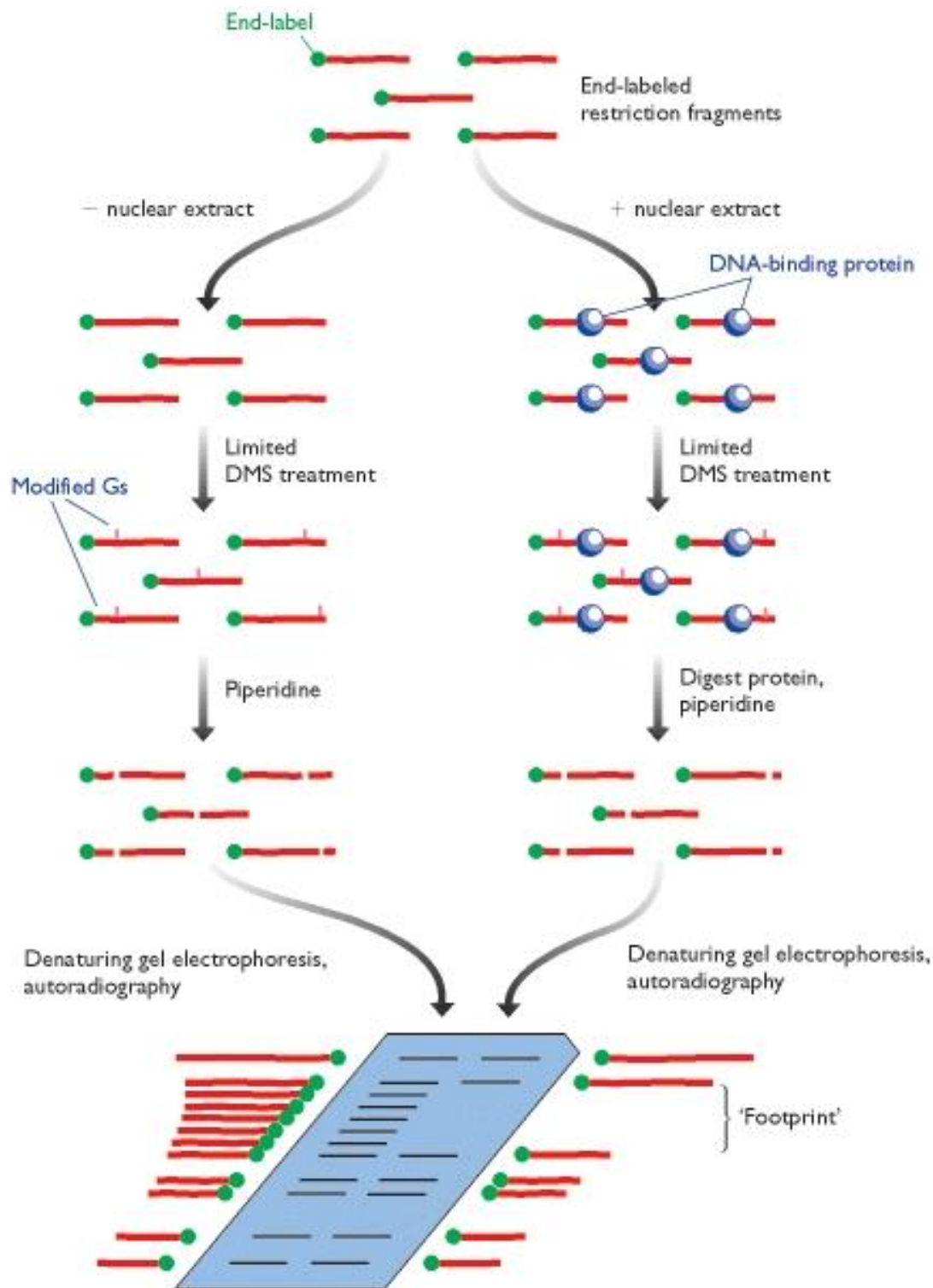


Figure 9.5 The dimethyl sulfate (DMS) modification protection assay. The technique is similar to DNase I footprinting (see Figure 9.4). Instead of DNase I digestion, the fragments are treated with limited amounts of DMS so that a single guanine base is methylated (more...)

Modification interference identifies nucleotides central to protein binding

Modification protection should not be confused with modification interference, a different technique with greater sensitivity in the study of protein binding (Hendrickson and Schleif, 1985). Modification interference works on the basis that if a nucleotide critical for protein binding is altered, for example by addition of a methyl group, then binding may be prevented. One of this family of techniques is illustrated in Figure 9.6. The DNA fragment, labeled at one end, is treated with the modification reagent, in this case dimethyl sulfate, under limiting conditions so that just one guanine per fragment is methylated. Now the binding protein or nuclear extract is added, and the fragments electrophoresed. Two bands are seen, one corresponding to the DNA-protein complex and one containing DNA without bound protein. The latter contains molecules that have been prevented from attaching to the protein because the methylation treatment has modified one or more Gs that are crucial for the binding. To identify which Gs are modified, the fragment is purified from the gel and treated with piperidine, a compound that cleaves DNA at methylguanine nucleotides. The result of this treatment is that each fragment is cut into two segments, one of which carries the label. The length(s) of the labeled fragment(s), determined by a second round of electrophoresis, tells us which nucleotide(s) in the original fragment were methylated and hence identifies the position in the DNA sequence of Gs that participate in the binding reaction. Equivalent techniques can be used to identify the A, C and T nucleotides involved in binding.

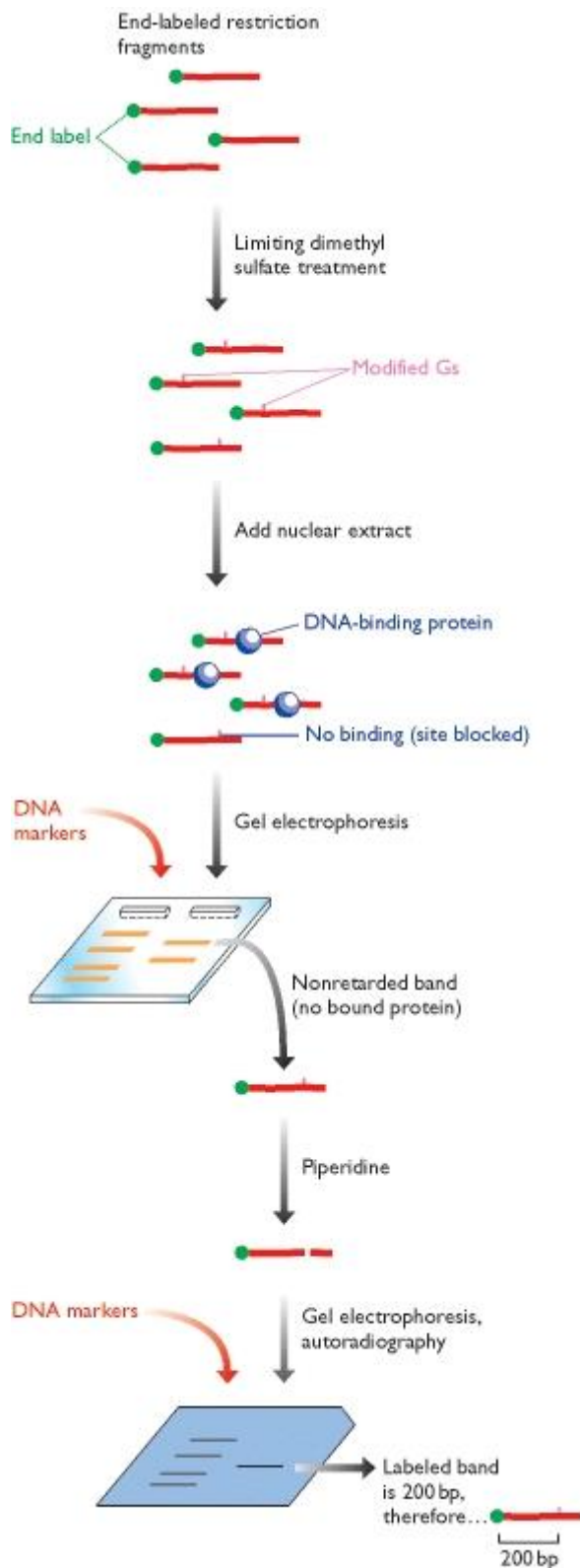


Figure 9.6

Dimethyl sulfate (DMS) modification interference assay. The method is described in the text. See the legend to Figure 9.4 for a description of the procedure used to obtain DNA fragments labeled at just one end.

9.1.2. Purifying a DNA-binding protein

Once a binding site has been identified in a DNA molecule, this sequence can be used to purify the DNA-binding protein, as a prelude to more detailed structural studies. The purification techniques utilize the ability of the protein to bind to its target site. One possibility is to use a form of affinity chromatography (Figure 9.7A). A DNA fragment or synthetic oligonucleotide that contains a protein binding site is immobilized in a chromatography column, usually by attaching one end of the DNA to a silica particle (Kadonaga, 1991). The protein extract is then passed through the column in a low-salt buffer, which promotes binding of proteins to their target sites. The binding protein specific for the immobilized sequence is retained in the column while all other proteins pass through. Once these unwanted proteins have been completely washed out, the column is eluted with a high-salt buffer, which destabilizes the DNA-protein complex. The pure binding protein can then be collected.

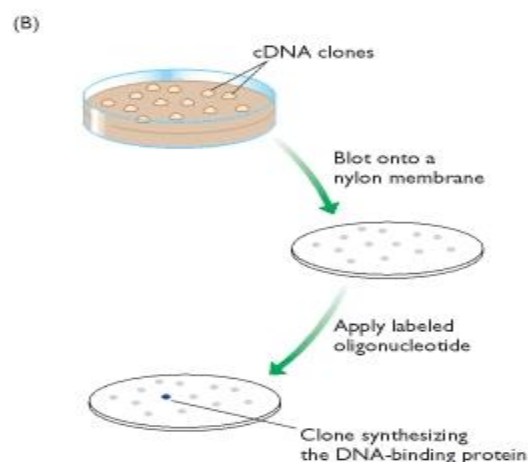
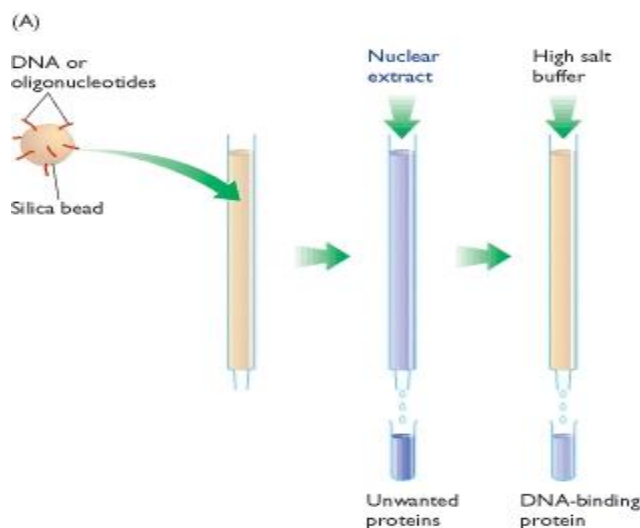


Figure 9.7

Two ways of purifying a DNA-binding protein. (A) Affinity chromatography. DNA fragments or synthetic oligonucleotides containing the attachment site for the binding protein are attached to silica beads and these packed into a chromatography column. The (more...)

An alternative is to screen a cloning library (Singh *et al.*, 1988). A library of cDNA clones, each synthesizing a different cloned protein from the organism being studied, is needed. These clones are blotted onto a nylon membrane in such a way that the protein content of each clone is retained (Figure 9.7B). The DNA fragment or oligonucleotide containing the protein binding site is labeled, and washed over the membrane. The DNA attaches to a blotted clone only if that clone has been synthesizing the appropriate DNA-binding protein. These clones are identified by detecting where the labeled DNA is located on the membrane. Samples of the clones can then be recovered from the master library and used to produce larger quantities of the binding protein.

9.1.3. Studying the structures of proteins and DNA-protein complexes

The availability of a pure sample of a DNA-binding protein makes possible the analysis of its structure, in isolation or attached to its DNA-binding site. This provides the most detailed information on the DNA-protein interaction, enabling the precise structure of the DNA-binding part of the protein to be determined, and allowing the identity and nature of the contacts with the DNA helix to be elucidated. Two techniques - X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy - are central to this area of research.

X-ray crystallography has broad applications in structure determination

X-ray crystallography is a long-established technique whose pedigree stretches back to the late 19th century. Indeed, Nobel prizes were awarded as early as 1915 to William and Lawrence Bragg, father and son, for working out the basic methodology and using it to determine the crystal structures of salts such as sodium chloride and zinc sulfide. The technique is based on X-ray diffraction. X-rays have very short wavelengths - between 0.01 and 10 nm - which is 4000 times shorter than visible light and comparable with the spacings between atoms in chemical structures. When a beam of X-rays is directed onto a crystal, some of the X-rays pass straight through, but others are diffracted and emerge from the crystal at a different angle from which they entered (Figure 9.8A). If the crystal is comprised of many copies of the same molecule, all positioned in a regular array, then different X-rays are diffracted in similar ways, resulting in overlapping circles of diffracted waves which interfere with one another. An X-ray-sensitive photographic film or electronic detector placed across the beam reveals a series of spots (Figure 9.8B), an X-ray diffraction pattern, from which the structure of the molecule in the crystal can be deduced.

The challenge with X-ray crystallography lies with the complexity of the methodology used to deduce the structure of a molecule from its diffraction pattern. The basic principles are that the relative positioning of the spots indicates the arrangement of the molecules in the crystal, and their relative intensities provide information on the structure of the molecule. The problem is that the more complex the molecule, the greater the number of spots and the larger the number of comparisons that must be made between them. Even with computational help the analysis is difficult and time consuming. If successful, the result is an electron density map (Figure 9.8C and D) which, with a protein, provides a chart of the folded polypeptide from which the positioning of structural features such as α -helices and β -sheets can be determined. If sufficiently detailed, the R groups of the individual amino acids in the polypeptide can be identified and their orientations relative to one another established, allowing deductions to be made about the hydrogen bonding and other chemical interactions occurring within the protein structure. With luck, these deductions lead to a detailed three-dimensional model of the protein (Rhodes, 1999).

The first protein structures to be determined by X-ray crystallography were for myoglobin and hemoglobin, resulting in further Nobel prizes, for Perutz and Kendrew in 1962. It still takes several months or longer to complete an X-ray crystallography analysis with a new protein, and there are many pitfalls that can prevent a successful conclusion being reached. In particular, it can often be difficult to obtain a suitable crystal of the protein. Despite these problems, the number of completed structures has gradually increased and now includes more than 50 DNA-binding proteins. An important innovation has been to crystallize DNA-binding proteins in the presence of their target sequences, the resulting protein-DNA structures revealing the precise positioning of the proteins relative to the double helix. It is from this type of information that most of our knowledge about the mode of action of DNA-binding proteins has been obtained.

NMR gives detailed structural information for small proteins

Like X-ray crystallography, NMR traces its origins to the early part of the 20th century, first being described in 1936 with the relevant Nobel prizes awarded in 1952. The principle of the technique is that rotation of a charged chemical nucleus generates a magnetic moment. When placed in an applied electromagnetic field, the spinning nucleus orientates in one of two ways, called α and β (Figure 9.9), the α -orientation (which is aligned with the magnetic field) having a slightly lower energy. In NMR spectroscopy the magnitude of this energy separation is determined by measuring the frequency of the electromagnetic radiation needed to induce the transition from α to β , the value being described as the resonance frequency of the nucleus

being studied. The critical point is that although each type of nucleus (e.g. ^1H , ^{13}C , ^{15}N) has its own specific resonance frequency, the measured frequency is often slightly different from the standard value (typically by less than 10 parts per million) because electrons in the vicinity of the rotating nucleus shield it to a certain extent from the applied magnetic field. This chemical shift (the difference between the observed resonance energy and the standard value for the nucleus being studied) enables the chemical environment of the nucleus to be inferred, and hence provides structural information. Particular types of analysis (called COSY and TOCSY) enable atoms linked by chemical bonds to the spinning nucleus to be identified; other analyses (e.g. NOESY) identify atoms that are close to the spinning nucleus in space but not directly connected to it.

Not all chemical nuclei are suitable for NMR. Most protein NMR projects are ^1H studies, the aim being to identify the chemical environments and covalent linkages of every hydrogen atom, and from this information to infer the overall structure of the protein. These studies are frequently supplemented by analyses of substituted proteins in which at least some of the carbon and/or nitrogen atoms have been replaced with the rare isotopes ^{13}C and ^{15}N , these also giving good results with NMR.

When successful, NMR results in the same level of resolution as X-ray crystallography and so provides very detailed information on protein structure (Evans, 1995). The main advantage of NMR is that it works with molecules in solution and so avoids the problems that sometimes occur when attempting to obtain crystals of a protein for X-ray analysis. Solution studies also offer greater flexibility if the aim is to examine changes in protein structure, for example during protein folding or in response to addition of a substrate. The disadvantage of NMR is that it is only suitable for relatively small proteins. There are several reasons for this, one being the need to identify the resonance frequencies for each, or as many as possible, of the ^1H or other nuclei being studied. This depends on the various nuclei having different chemical shifts so that their frequencies do not overlap. The larger the protein, the greater the number of nuclei and the greater the chances that frequencies overlap and structural information is lost. Although this limits the applicability of NMR, the technique is still very valuable. There are many interesting proteins that are small enough to be studied by NMR, and important information can also be obtained by structural analysis of peptides which, although not complete proteins, can act as models for aspects of protein activity such as nucleic acid binding.

9.1.4. The special features of DNA-binding proteins

Now that we have examined the methods used to study DNA-binding proteins, we can turn our attention to the proteins themselves. Our main interest lies with those proteins that are able to target a specific nucleotide sequence and hence bind to a limited number of positions on a DNA molecule, this being the type of interaction that is most important in expression of the genome. To bind in this specific fashion a protein must make contact with the double helix in such a way that the nucleotide sequence can be recognized, which generally requires that part of the protein penetrates into the major and/or minor grooves of the helix (see Figures 1.11A and 1.12) in order to achieve direct readout of the sequence (Section 9.1.5). This is usually accompanied by more general interactions with the surface of the molecule, which may simply stabilize the DNA-protein complex or which may be aimed at accessing indirect information on nucleotide sequence that is provided by the conformation of the helix.

When the structures of sequence-specific DNA-binding proteins are compared, it is immediately evident that the family as a whole can be divided into a limited number of different groups on the basis of the structure of the segment of the protein that interacts with the DNA molecule (Table 9.2; Luisi, 1995). Each of these DNA-binding motifs is present in a range of proteins, often from very different organisms, and at least some of them probably evolved more than once. We will look at two in detail - the **helix-turn-helix** (HTH) motif and the zinc finger - and then briefly survey the others.

Motif	Examples of proteins with this motif
Sequence-specific DNA-binding motifs	
Helix-turn-helix family	
Standard helix-turn-helix	<i>Escherichia coli</i> lactose repressor, tryptophan repressor
Homeodomain	<i>Drosophila</i> Antennapedia protein
Paired homeodomain	Vertebrate Pax transcription factors
POU domain	Vertebrate regulatory proteins PIT-1, OCT-1 and OCT-2
Winged helix-turn-helix	GABP regulatory protein of higher eukaryotes

Motif	Examples of proteins with this motif
High mobility group (HMG) domain	Mammalian sex determination protein SRY
Zinc-finger family	
Cys ₂ His ₂ finger	Transcription factor TFIIA of eukaryotes
Multi-cysteine zinc finger	Steroid receptor family of higher eukaryotes
Zinc binuclear cluster	Yeast GAL4 transcription factor
Basic domain	Yeast GCN4 transcription factor
Ribbon-helix-helix	Bacterial MetJ, Arc and Mnt repressors
TBP domain	Eukaryotic TATA-binding protein
β-Barrel dimer	Papillomavirus E2 protein
Rel homology domain (RHB)	Mammalian transcription factor NF-κB
Non-specific DNA-binding motifs	
Histone fold	Eukaryotic histones
HU/IHF motif ^a	Bacterial HU and IHF proteins
Polymerase cleft	DNA and RNA polymerases

Table 9.2

DNA-binding motifs.

The helix-turn-helix motif is present in prokaryotic and eukaryotic proteins

The HTH motif was the first DNA-binding structure to be identified (Harrison and Aggarwal, 1990). As the name suggests, the motif is made up of two α -helices separated by a turn (Figure 9.10). The latter is not a random conformation but a specific structure, referred to as a **β -turn**, made up of four amino acids, the second of which is usually glycine. This turn, in conjunction with the first α -helix, positions the second α -helix on the surface of the protein in an orientation that enables it to fit inside the major groove of a DNA molecule. This second α -helix is therefore the recognition helix that makes the vital contacts which enable the DNA sequence to be read. The HTH structure is usually 20 or so amino acids in length and so is just a small part of the protein as a whole. Some of the other parts of the protein form attachments with the surface of the DNA molecule, primarily to aid the correct positioning of the recognition helix within the major groove.

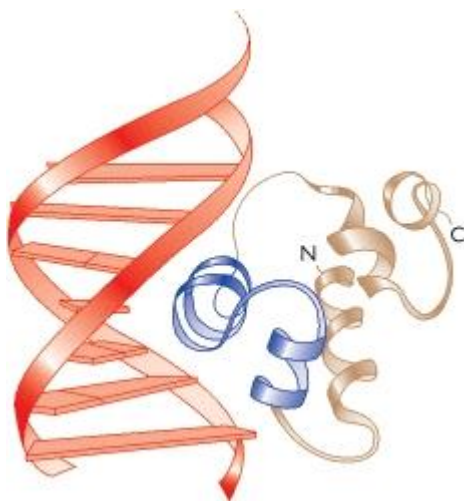


Figure 9.10

The helix-turn-helix motif. The drawing shows the orientation of the helix-turn-helix motif (in blue) of the *Escherichia coli* bacteriophage 434 repressor in the major groove of the DNA double helix. 'N' and 'C' indicate (more...)

Many prokaryotic and eukaryotic DNA-binding proteins utilize an HTH motif. In bacteria, HTH motifs are present in some of the best studied regulatory proteins, which switch on and off the expression of individual genes. An example is the lactose repressor, which regulates expression of the lactose operon (Sections 2.3.2 and 9.3.1). The various eukaryotic HTH proteins include many whose DNA-binding properties are important in the developmental regulation of genome expression, such as the homeodomain proteins, whose roles we will examine in Section 12.3.3. The homeodomain is an extended HTH motif possessed by each

of these proteins. It is made up of 60 amino acids which form four α -helices, numbers 2 and 3 separated by a β -turn, with number 3 acting as the recognition helix and number 1 making contacts within the minor groove (Figure 9.11). Other versions of the HTH motif found in eukaryotes include:

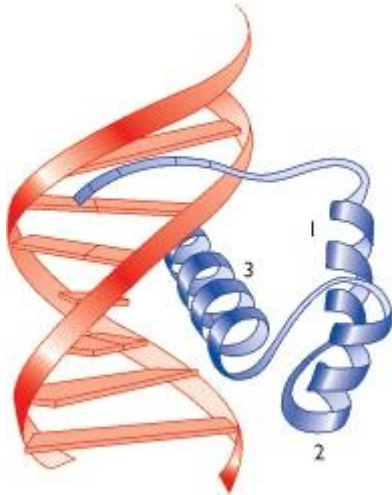


Figure 9.11

The homeodomain motif. The first three helices of a typical homeodomain are shown with helix 3 orientated in the major groove and helix 1 making contacts in the minor groove. Helices 1–3 run in the N→C direction along the motif. Reprinted (more...)

- The POU domain, which is usually found in proteins that also have a homeodomain, the two motifs probably working together by binding different regions of a double helix. The name ‘POU’ comes from the initial letters of the names of the first proteins found to contain this motif (Herr *et al.*, 1988).
- The **winged helix-turn-helix** motif, which is another extended version of the basic HTH structure, this one with a third α -helix on one side of the HTH motif and a β -sheet on the other side.

Many proteins, prokaryotic and eukaryotic, possess an HTH motif, but the details of the interaction of the recognition helix with the major groove are not exactly the same in all cases. The length of the recognition helix varies, generally being longer in eukaryotic proteins, the orientation of the helix in the major groove is not always the same, and the position within the recognition helix of those amino acids that make contacts with nucleotides is different.

Zinc fingers are common in eukaryotic proteins

The second type of DNA-binding motif that we will look at in detail is the zinc finger, which is rare in prokaryotic proteins but very common in eukaryotes (Mackay and Crossley, 1998).

There appear to be more than 500 different zinc-finger proteins in the worm *Caenorhabditis elegans*, out of a total 19 000 proteins (Clarke and Berg, 1998), and it is estimated that 1% of all mammalian genes code for zinc-finger proteins.

There are at least six different versions of the zinc finger. The first to be studied in detail was the **Cys₂His₂ finger**, which comprises a series of 12 or so amino acids, including two cysteines and two histidines, which form a segment of β -sheet followed by an α -helix. These two structures, which form the 'finger' projecting from the surface of the protein, hold between them a bound zinc atom, coordinated to the two cysteines and two histidines (Figure 9.12). The α -helix is the part of the motif that makes the critical contacts within the major groove, its positioning within the groove being determined by the β -sheet, which interacts with the sugar-phosphate backbone of the DNA, and the zinc atom, which holds the sheet and helix in the appropriate positions relative to one another. Other versions of the zinc finger differ in the structure of the finger, some lacking the sheet component and consisting simply of one or more α -helices, and the precise way in which the zinc atom is held in place also varies. For example, the multicysteine zinc fingers lack histidines, the zinc atom being coordinated between four cysteines.

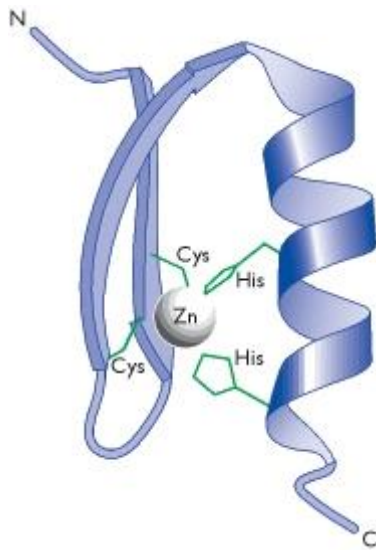


Figure 9.12

The Cys₂His₂ zinc finger. This particular zinc finger is from the yeast SWI5 protein. The zinc atom is held between two cysteines within the β -sheet of the motif and two histidines in the α -helix. The solid green lines indicate the R groups (more...)

An interesting feature of the zinc finger is that multiple copies of the finger are sometimes found on a single protein. Several have two, three or four fingers, but there are examples with many more than this - 37 for one toad protein. In most cases, the individual zinc fingers are

thought to make independent contacts with the DNA molecule, but in some cases the relationship between different fingers is more complex. In one particular group of proteins - the nuclear or steroid receptor family - two α -helices containing six cysteines combine to coordinate two zinc atoms in a single DNA-binding domain, larger than a standard zinc finger (Figure 9.13). Within this motif it appears that one of the α -helices enters the major groove whereas the second makes contacts with other proteins.

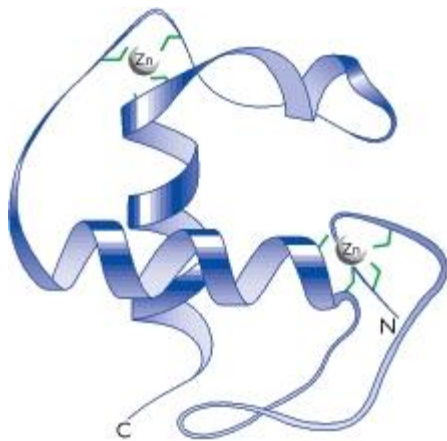


Figure 9.13

The steroid receptor zinc finger. The R groups of the amino acids involved in the interactions with the zinc atoms are shown as solid green lines. 'N' and 'C' indicate the N- and C-termini of the motif, respectively. Reprinted (more...)

RNA-binding proteins also have specific motifs that form the attachment with the RNA molecule. The most important of these are as follows:

- The ribonucleoprotein (RNP) domain comprises four β -strands and two α -helices in the order β - α - β - α - β . The two central β -strands make the critical attachments with the RNA molecule. The RNP domain is the commonest RNA-binding motif and has been found in more than 250 proteins.
- The **double-stranded RNA binding domain (dsRBD)** is similar to the RNP domain but with the structure α - β - β - α . The RNA-binding function lies between the β and α at the end of the structure. As the name implies, the motif is found in proteins that bind double-stranded RNA (Fierro-Monti and Mathews, 2000).
- The **κ -homology domain** has the structure β - α - α - β - α , with the binding function between the pair of α -helices. It is relatively uncommon but present in at least one nuclear RNA-binding protein.
- The DNA-binding homeodomain may also have RNA-binding activity in some proteins. One ribosomal protein uses a structure similar to a homeodomain to attach

to rRNA, and some homeodomain proteins such as Bicoid of *Drosophila melanogaster* (Section 12.3.3) can bind both DNA and RNA.

Box 9.1

RNA-binding motifs. RNA-binding proteins also have specific motifs that form the attachment with the RNA molecule. The most important of these are as follows: The ribonucleoprotein (RNP) domain comprises four β -strands and two α -helices(more...)

Other DNA-binding motifs

The various other DNA-binding motifs that have been discovered in different proteins include:

- The basic domain, in which the DNA recognition structure is an α -helix that contains a high number of basic amino acids (e.g. arginine, serine and threonine). A peculiarity of this motif is that the α -helix only forms when the protein interacts with DNA: in the unbound state the helix has a disorganized structure. Basic domains are found in a number of eukaryotic proteins involved in transcription of DNA into RNA.
- The **ribbon-helix-helix** motif, which is one of the few motifs that achieves sequence-specific DNA binding without making use of an α -helix as the recognition structure. Instead, the ribbon (i.e. two strands of a β -sheet) makes contact with the major groove (Figure 9.14). Ribbon-helix-helix motifs are found in some gene-regulatory proteins in bacteria.
- The TBP domain has so far only been discovered in the TATA-binding protein (Section 9.2.3), after which it is named (Kim *et al.*, 1993). As with the ribbon-helix-helix motif, the recognition structure is a β -sheet, but in this case the main contacts are with the minor, not major, groove of the DNA molecule.

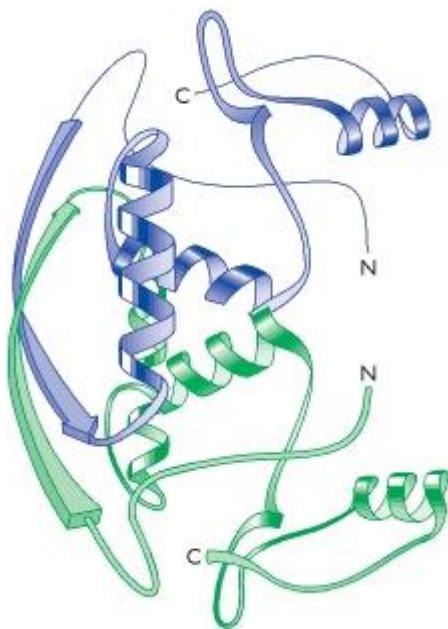


Figure 9.14

The ribbon-helix-helix motif. The drawing is of the ribbon-helix-helix motif of the *Escherichia coli* MetJ repressor, which consists of a dimer of two identical proteins, one shown in blue and the other in green. The β -strands at the left of the (more...)

9.1.5. The interaction between DNA and its binding proteins

In recent years our understanding of the part played by the DNA molecule in the interaction with a binding protein has begun to change. It has always been accepted that proteins that recognize a specific sequence as their binding site can locate this site by forming contacts with chemical groups attached to the nitrogenous bases that are exposed within the major and minor grooves that spiral around the double helix (see Figure 1.11A). It is now recognized that the nucleotide sequence also influences the precise conformation of each region of the helix, and that these conformational features represent a second, less direct way in which the DNA sequence can influence protein binding.

Direct readout of the nucleotide sequence

It was clear from the double helix structure described by Watson and Crick (Section 1.1.3) that although the nucleotide bases are on the inside of the DNA molecule, they are not entirely buried, and some of the chemical groups attached to the purine and pyrimidine bases are accessible from outside the helix. Direct readout of the nucleotide sequence should therefore be possible without breaking the base pairs and opening up the molecule.

In order to form chemical bonds with groups attached to the nucleotide bases, a binding protein must make contacts within one or both of the grooves on the surface of the helix. With the B-form of DNA, the identity and orientation of the exposed parts of the bases within the major groove is such that most sequences can be read unambiguously, whereas within the minor groove it is possible to identify if each base pair is A-T or G-C but difficult to know which nucleotide of the pair is in which strand of the helix (Figure 9.15; Kielkopf *et al.*, 1998). Direct readout of the B-form therefore predominantly involves contacts in the major groove. With other DNA types there is much less information on the contacts formed with binding proteins, but the picture is likely to be quite different. In the A-form, for example, the major groove is deep and narrow and less easily penetrated by any part of a protein molecule (see Table 1.1). The shallower minor groove is therefore likely to play the main part in direct readout. With Z-DNA, the major groove is virtually non-existent and direct readout is possible to a certain extent without moving beyond the surface of the helix.

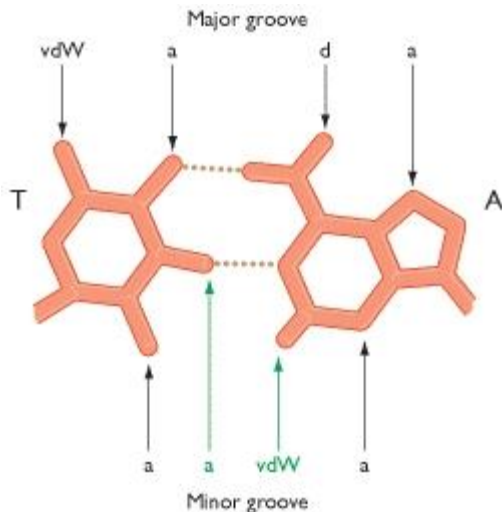


Figure 9.15

Recognition of an A-T base pair in the B-form double helix. An A-T base pair is shown in outline (see Figure 1.11B), with arrows indicating the chemical features that can be recognized by accessing the base pair via the major groove (above) and minor (more...)

The nucleotide sequence has a number of indirect effects on helix structure

The recent change in our view of DNA structure concerns the influence of the nucleotide sequence on the conformation of the helix at different positions along its length. Originally it was thought that cellular DNA molecules have fairly uniform structures, made up mainly of the B-form of the double helix. Some short segments might be in the A-form, and there might be some Z-DNA tracts, especially near the ends of a molecule, but the majority of the length of a double helix would be unvarying B-DNA. We now recognize that DNA is much more polymorphic, and that it is possible for the A-, B- and Z-DNA configurations, and intermediates between them, to coexist within a single DNA molecule, different parts of the molecule having different structures. These conformational variations are sequence dependent, being largely the result of the base-stacking interactions that occur between adjacent base pairs. As well as being responsible, along with base-pairing, for the stability of helix, the base-stacking also influences the amount of rotation that occurs around the covalent bonds within individual nucleotides and hence determines the conformation of the helix at a particular position. The rotational possibilities in one base pair are influenced, via the base-stacking interactions, by the identities of the neighboring base pairs. This means that the nucleotide sequence indirectly affects the overall conformation of the helix, possibly providing structural information that a binding protein can use to help it locate its appropriate attachment site on a DNA molecule. At present this is just a theoretical possibility as no protein that specifically recognizes a non-B form of the helix has been identified, but many

researchers believe that helix conformation is likely to play some role in the interaction between DNA and protein.

A second type of conformational change is DNA bending (Travers, 1995). This does not refer to the natural flexibility of DNA which enables it to form circles and supercoils, but instead to localized positions where the nucleotide sequence causes the DNA to bend. Like other conformational variations, DNA bending is sequence dependent. In particular, a DNA molecule in which one polynucleotide contains two or more groups of repeated adenines, each group comprising 3–5 As, with individual groups separated by 10 or 11 nucleotides, will bend at the 3' end of the adenine-rich region (Young and Beveridge, 1998). As with helix conformation, it is not yet known to what extent DNA bending influences protein binding, although protein-induced bending at flexible sites has a clearly demonstrated function in the regulation of some genes (e.g. Falvo *et al.*, 1995; Section 9.3.2).

Contacts between DNA and proteins

The contacts formed between DNA and its binding proteins are non-covalent. Within the major groove, hydrogen bonds form between the nucleotide bases and the R groups of amino acids in the recognition structure of the protein, whereas in the minor groove hydrophobic interactions are more important. On the surface of the helix, the major interactions are electrostatic, between the negative charges on the phosphate component of each nucleotide and the positive charges on the R groups of amino acids such as lysine and arginine, although some hydrogen bonding also occurs. In some cases, hydrogen bonding on the surface of the helix or in the major groove is direct between DNA and protein; in others it is mediated by water molecules. Few generalizations can be made: at this level of DNA-protein interaction each example has its own unique features and the details of the bonding have to be worked out by structural studies rather than by comparisons with other proteins.

Most proteins that recognize specific sequences are also able to bind non-specifically to other parts of a DNA molecule. In fact it has been suggested that the amount of DNA in a cell is so large, and the numbers of each binding protein so small, that the proteins spend most, if not all, of their time attached non-specifically to DNA (Stormo and Fields, 1998). The distinction between the non-specific and specific forms of binding is that the latter is more favorable in thermodynamic terms. As a result, a protein is able to bind to its specific site even though there are literally millions of other sites to which it could attach non-specifically. To achieve this thermodynamic favorability, the specific binding process must involve the greatest possible number of DNA-protein contacts, which explains in part why the recognition structures of many DNA-binding motifs have evolved to fit snugly into the major groove of

the helix, where the opportunity for DNA-protein contacts is greatest. It also explains why some DNA-protein interactions result in conformational changes to one or other partner, increasing still further the complementarity of the interacting surfaces, and allowing additional bonding to occur.

The need to maximize contacts in order to ensure specificity is also one of the reasons why many DNA-binding proteins are dimers, consisting of two proteins attached to one another. This is the case for most HTH proteins and many of the zinc-finger type. Dimerization occurs in such a way that the DNA-binding motifs of the two proteins are both able to access the helix, possibly with some degree of cooperativity between them, so that the resulting number of contacts is greater than twice the number achievable by a monomer. As well as their DNA-binding motifs, many proteins contain additional characteristic domains that participate in the protein-protein contacts that result in dimer formation. One of these is the leucine zipper, which is an α -helix that coils more tightly than normal and presents a series of leucines on one of its faces. These can form contacts with the leucines of the zipper on a second protein, forming the dimer (Figure 9.16). A second dimerization domain is, rather unfortunately, called the **helix-loop-helix** motif, which is distinct from, and should not be confused with, the helix-turn-helix DNA-binding motif.

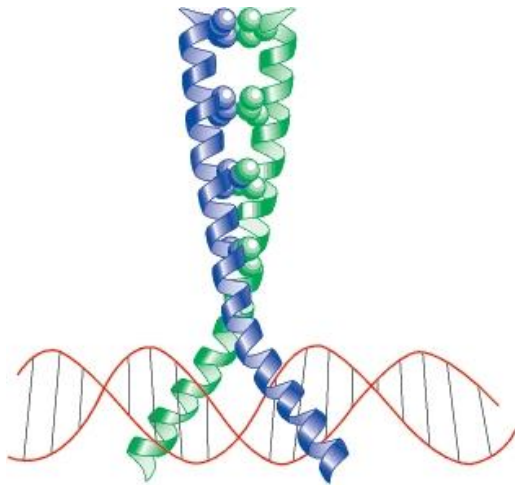


Figure 9.16

A leucine zipper. This is a bZIP type of leucine zipper. The blue and green structures are parts of different proteins. Each set of spheres represents the R-group of a leucine amino acid. Leucines in the two helices associate with one another via hydrophobic (more...)

Determining the Function of Genes during Development

Transgenic cells and organisms

While it is important to know the sequence of a gene and its temporal-spatial pattern of expression, what's really crucial is to know the functions of that gene during development. Recently developed techniques have enabled us to study gene function by moving certain genes into and out of embryonic cells.

WEBSITE

4.6 Bioinformatics. Information about gene regulation and developmental pathways may soon be modeled on computers. Accessibility to this information may enable researchers to design experiments that have higher chances of success. <http://www.devbio.com/chap04/link0406.shtml>

Inserting new DNA into a cell

Cloned pieces of DNA can be isolated, modified (if so desired), and placed into cells by several means. One very direct technique is **microinjection**, in which a solution containing the cloned gene is injected very carefully into the nucleus of a cell (Capecchi 1980). This is an especially useful technique for injecting genes into newly fertilized eggs, since the haploid nuclei of the sperm and egg are relatively large (Figure 4.18). In **transfection**, DNA is incorporated directly into cells by incubating them in a solution that makes them “drink” it in. The chances of a DNA fragment being incorporated into the chromosomes in this way are relatively small, however, so the DNA of interest is usually mixed with another gene, such as a gene encoding resistance to a particular antibiotic, that enables those rare cells that incorporate the DNA to survive under culture conditions that will kill all the other cells (Perucho et al. 1980; Robins et al. 1981). Another technique is **electroporation**, in which a high-voltage pulse “pushes” the DNA into the cells.



Figure 4.18

Insertion of new DNA into embryonic cells. Here, DNA (from cloned genes) is injected into the pronucleus of a mouse egg. (From Wagner et al. 1981; photograph courtesy of T. E. Wagner.)

A more “natural” way of getting genes into cells is to put the cloned gene into a **transposable element** or **retroviral vector**. These are naturally occurring mobile regions of DNA that can

integrate themselves into the genome of an organism. Retroviruses are RNA-containing viruses. Within a host cell, they make a DNA copy of themselves (using their own virally encoded reverse transcriptase); the copy then becomes double-stranded and integrates itself into a host chromosome. The integration is accomplished by two identical sequences (long terminal repeats) at the ends of the retroviral DNA. Retroviral vectors are made by removing the viral packaging genes (needed for the exit of viruses from the cell) from the center of a mouse retrovirus. This extraction creates a vacant site where other genes can be placed. By using the appropriate restriction enzymes, researchers can insert an isolated gene (such as a gene isolated by PCR) and insert it into a retroviral vector. These retroviral vectors infect mouse cells with an efficiency approaching 100%. Similarly, in *Drosophila*, new genes can be carried into a fly via **P elements**. These DNA sequences are naturally occurring transposable elements that can integrate like viruses into any region of the *Drosophila* genome. Moreover, they can be isolated, and cloned genes can be inserted into the center of the P element. When the recombined P element is injected into a *Drosophila* oocyte, it can integrate into the DNA and provide the embryo with the new gene (Spradling and Rubin 1982).

Chimeric mice

The techniques described above have been used to transfer genes into every cell of the mouse embryo (Figure 4.19). During early mouse development, there is a stage when only two cell types are present: the outer trophoblast cells, which will form the fetal portion of the placenta, and the inner cell mass, whose cells will give rise to the embryo itself. These inner cells are the cells whose separation can lead to twins (Chapters 3 and 11), and if an inner cell mass blastomere of one mouse is transferred into the embryo of a second mouse, that donor cell can contribute to every organ of the host embryo.

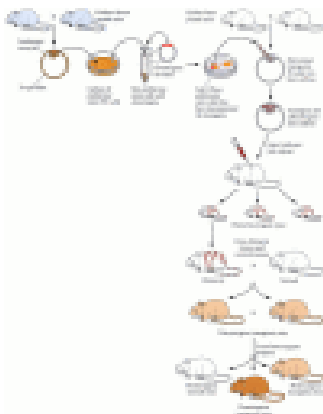


Figure 4.19

Production of transgenic mice. Embryonic stem cells from one mouse are cultured and their genome altered by the addition of a cloned gene. These transgenic cells are selected and then injected into the early stages of a host mouse embryo. Here, the transgenic (more...)

Inner cell mass blastomeres can be isolated from the embryo and cultured in vitro; such cultured cells are called **embryonic stem cells (ES cells)**. ES cells retain their totipotency, and each of them can contribute to all organs if injected into a host embryo (Gardner 1968; Moustafa and Brinster 1972). Moreover, once in culture, these cells can be treated as described in the preceding section so that they will incorporate new DNA. A treated ES cell (the entire cell, not just the DNA) can then be injected into another early-stage mouse embryo, and will integrate into the host embryo. The result is a **chimeric mouse**.^{*} Some of this mouse's cells will be derived from the host embryonic stem cells, but some portion of its cells will be derived from the treated embryonic stem cell. If the treated cells become part of the germ line of the mouse, some of its gametes will be derived from the donor cell. If such a chimeric mouse is mated with a wild-type mouse, some of its progeny will carry one copy of the inserted gene. When these heterozygous progeny are mated to one another, about 25% of the resulting offspring will carry two copies of the inserted gene in every cell of their bodies (Gossler et al. 1986). Thus, in three generations—the chimeric mouse, the heterozygous mouse, and the homozygous mouse—a gene that was cloned from some other organism will be present in both copies of the chromosomes within the mouse genome. Strains of such transgenic mice have been particularly useful in determining how genes are regulated during development.

Gene targeting (“knockout”) experiments

The analysis of early mammalian embryos has long been hindered by our inability to breed and select animals with mutations that affect early embryonic development. This block has been circumvented by the techniques of **gene targeting** (or, as it is sometimes called, **gene knockout**). These techniques are similar to those that generate transgenic mice, but instead of *adding* genes, gene targeting *replaces* wild-type alleles with mutant ones. As an example, we will look at the gene knockout of bone morphogenetic protein 7 (BMP7). Bone morphogenetic proteins are involved in numerous developmental interactions whereby one set of cells interacts with other neighboring cells to alter their properties. BMP7 has been implicated as a protein that prevents cell death and promotes cell division in several developing organs. Dudley and his colleagues 1995 used gene targeting to find the function of BMP7 in the development of the mouse. First, they isolated the *BMP7* gene, cut it at one

site with a restriction enzyme, and inserted a bacterial gene for neomycin resistance into that site (Figure 4.20). In other words, they mutated the *BMP7* gene by inserting into it a large piece of foreign DNA, destroying the ability of the BMP7 protein to function. These mutant *BMP7* genes were electroporated into ES cells that were sensitive to neomycin. Once inside the nucleus of an ES cell, the mutated *BMP7* gene may replace a normal allele of *BMP7* by a process called homologous recombination. In this process, the enzymes involved in DNA repair and replication incorporate the mutant gene in the place of the normal copy. It's a rare event, but such cells can be selected by growing the ES cells in neomycin. Most of the cells are killed by the drug, but the ones that have acquired resistance from the incorporated gene survive. The resulting cells have one normal *BMP7* gene and one mutated *BMP7* gene. These heterozygous ES cells were then microinjected into mouse blastocysts, where they were integrated into the cells of the embryo. The resulting mice were chimeras composed of wild-type cells from the host embryo and heterozygous *BMP7*-containing cells from the donor ES cells. The chimeras were mated to wild-type mice, producing progeny that were heterozygous for the *BMP7* gene. These heterozygous mice were then bred with each other, and about 25% of their progeny carried two copies of the mutated *BMP7* gene. These homozygous mutant mice lacked eyes and kidneys (Figure 4.21). In the absence of BMP7, it appears that many of the cells that normally form these two organs stop dividing and die. In this way, gene targeting can be used to analyze the roles of particular genes during mammalian development.



Figure 4.20

Technique for gene targeting. In this case, the targeted gene is *BMP7*. (A) Embryonic stem (ES) cells from a mouse blastocyst are cultured. (B) Cloned *BMP7* genes are cut with a restriction enzyme, and a neomycin resistance gene is inserted into the region (more...)

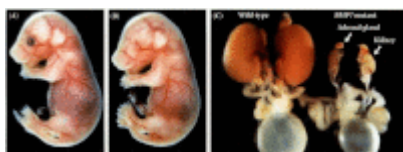


Figure 4.21

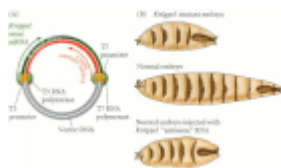
Morphological analysis of *BMP7* knockout mice. A wild-type (A) and a homozygous *BMP7*-deficient mouse (B) at day 17 of their 21-day gestation. The *BMP7*-deficient mouse lacks eyes. The kidneys of these mice at day 19 of gestation are shown in (C). The kidney (more...)

Go to:

Determining the function of a message: Antisense RNA

Another method for determining the function of a gene is to use “antisense” copies of its message to block the function of that message. Antisense RNA allows developmental biologists to determine the function of genes during development and to analyze the action of genes that would otherwise be inaccessible for genetic analysis.

Antisense messages can be generated by cloning DNA into vectors that have promoters at both ends of the inserted gene. When incubated with a particular RNA polymerase and nucleotide triphosphates, the promoter will initiate transcription of the message “in the wrong direction.” In so doing, it synthesizes a transcript that is complementary to the natural one (Figure 4.23A). This complementary transcript is called **antisense RNA** because it is the complement of the original (“sense”) message. When large amounts of antisense RNA are injected or transfected into cells containing the normal mRNA from the same gene, the antisense RNA binds to the normal message, and the resulting double-stranded nucleic acid is degraded. (Cells have enzymes to digest double-stranded nucleic acids in the cytoplasm.) This causes a functional deletion of the message, just as if there were a deletion mutation for that gene.

**Figure 4.23**

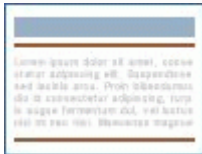
Use of antisense RNA to examine the roles of genes in development. (A) An antisense message (in this case, to the *Krüppel* gene of *Drosophila*) is produced by placing the cloned cDNA fragment for the *Krüppel* message between two strong promoters (more...)

The similarities between the phenotypes produced by a loss-of-function mutation and by antisense RNA treatment were seen when antisense RNA was made to the *Krüppel* gene of *Drosophila*. *Krüppel* is critical for forming the thorax and abdomen of the fly. If this gene is absent, fly larvae die because they lack thoracic and anterior abdominal segments (Figure

4.23B). A similar situation is created when large amounts of antisense RNA against the *Krüppel* message are injected into early fly embryos (Rosenberg et al. 1985).

WEBSITE

4.8 RNA interference. Soaking the nematode *C. elegans* in a solution containing double-stranded RNA will knock out the expression of that gene not only in the soaked animal, but also in the progeny of that animal. <http://www.devbio.com/chap04/link0408.shtml>



Box

Human Somatic and Germ Line Gene Therapy. *Embryonic Stem Cells* In 1998, two laboratories (Gearhart 1998; Thomson et al. 1998) announced that they had derived human embryonic stem cells. In some instances, these cells were derived from inner cell masses (more...)

Go to:

Footnotes

*

It is critical to note the difference between a chimera and a hybrid. A **hybrid** results from the union of two different genomes within the same cell: the offspring of an AA genotype parent and an aa genotype parent is an Aa hybrid. A **chimera** results when cells of different genetic constitution appear in the same organism. The term is apt: it refers to a mythical beast with a lion's head, a goat's body, and a serpent's tail.

Gene targeting (also, replacement strategy based on homologous recombination) is a genetic technique that uses homologous recombination to change an endogenous gene. The method can be used to delete a gene, remove exons, add a gene, and introduce point mutations. Gene targeting can be permanent or conditional. Conditions can be a specific time during development / life of the organism or limitation to a specific tissue, for example. Gene targeting requires the creation of a specific vector for each gene of interest. However, it can be used for any gene, regardless of transcriptional activity or gene size.

Methods[edit]

Gene targeting methods are established for several model organisms and may vary depending on the species used. In general, a targeting construct made out of DNA is generated

in bacteria. It typically contains part of the gene to be targeted, a reporter gene, and a (dominant) selectable marker.

To target genes in mice, this construct is then inserted into mouse embryonic stem cells in culture. After cells with the correct insertion have been selected, they can be used to contribute to a mouse's tissue via embryo injection. Finally, chimeric mice where the modified cells made up the reproductive organs are selected for via breeding. After this step the entire body of the mouse is based on the previously selected embryonic stem cell.

To target genes in moss, this construct is incubated together with freshly isolated protoplasts and with Polyethylene glycol. As mosses are haploid organisms,^[2] regenerating moss filaments (protonema) can directly be screened for gene targeting, either by treatment with antibiotics or with PCR. Unique among plants, this procedure for reverse genetics is as efficient as in yeast.^[3] Using modified procedures, gene targeting has also been successfully applied to cattle, sheep, swine, and many fungi.

The frequency of gene targeting can be significantly enhanced through the use of engineered endonucleases such as zinc finger nucleases,^[4] engineered homing endonucleases,^[5] and nucleases based on engineered TAL effectors.^[6] To date, this method has been applied to a number of species including *Drosophila melanogaster*,^[4] tobacco,^{[7][8]} corn,^[9] human cells,^[10] mice,^[11] and rats.^[11]

Comparison with gene trapping

Gene trapping is based on random insertion of a cassette while gene targeting targets a specific gene. Cassettes can be used for many different things while the flanking homology regions of gene targeting cassettes need to be adapted for each gene. This makes gene trapping more easily amenable for large scale projects than targeting. On the other hand, gene targeting can be used for genes with low transcriptions that would go undetected in a trap screen. Also, the probability of trapping increases with intron size. For gene targeting these compact genes are just as easily altered.

Applications

Gene targeting has been widely used to study human genetic diseases by removing ("knocking out"), or adding ("knocking in"), specific mutations of interest to a variety of models. Previously used to engineer rat cell models, advances in gene targeting technologies are enabling the creation of a new wave of isogenic human disease models. These models are the most accurate in-vitro models available to researchers to date, and are facilitating the development of new personalized drugs and diagnostics, particularly in the field of cancer.^[12]

2007 Nobel prize

Mario R. Capecchi, Martin J. Evans and Oliver Smithies were declared laureates of the 2007 Nobel Prize in Physiology or Medicine for their work on "principles for introducing specific gene modifications in mice by the use of embryonic stem cells", or gene targeting

Gene knock out

A **gene knockout** (abbreviation: **KO**) is a genetic technique in which one of an organism's genes is made inoperative ("knocked out" of the organism). Also known as **knockout organisms** or simply **knockouts**, they are used in learning about a gene that has been sequenced, but which has an unknown or incompletely known function. Researchers draw inferences from the difference between the knockout organism and normal individuals. The term also refers to the process of creating such an organism, as in "knocking out" a gene. The technique is essentially the opposite of a gene knockin. Knocking out two genes simultaneously in an organism is known as a **double knockout (DKO)**. Similarly the terms **triple knockout (TKO)** and **quadruple knockouts (QKO)** are used to describe three or four knocked out genes, respectively.

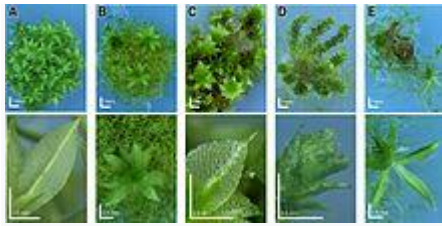
Method



A laboratory mouse in which a gene affecting hair growth has been knocked out (left), is shown next to a normal lab mouse.

Knockout is accomplished through a combination of techniques, beginning in the test tube with a plasmid, a bacterial artificial chromosome or other DNA construct, and proceeding to cell culture. Individual cells are genetically transfected with the DNA construct. Often the goal is to create a transgenic animal that has the altered gene. If so, embryonic stem cells are genetically transformed and inserted into early embryos. Resulting animals with the genetic change in their germline cells can then often pass the gene knockout to future generations.

To create knockout moss, transfection of protoplasts is the preferred method. Such transformed *Physcomitrella*-protoplasts directly regenerate into fertile moss plants. Eight weeks after transfection, the plants can be screened for gene targeting via PCR.^[1]



Wild-type Physcomitrella and knockout mosses: Deviating phenotypes induced in gene-disruption library transformants. *Physcomitrella* wild-type and transformed plants were grown on minimal Knop medium to induce differentiation and development of gametophores. For each plant, an overview (upper row; scale bar corresponds to 1 mm) and a close-up (bottom row; scale bar equals 0.5 mm) are shown. A: Haploid wild-type moss plant completely covered with leafy gametophores and close-up of wild-type leaf. B–D: Different mutants.^[2]

The construct is engineered to recombine with the target gene, which is accomplished by incorporating sequences from the gene itself into the construct. Recombination then occurs in the region of that sequence within the gene, resulting in the insertion of a foreign sequence to disrupt the gene. With its sequence interrupted, the altered gene in most cases will be translated into a nonfunctional protein, if it is translated at all.



A knockout mouse (left) that is a model of obesity, compared with a normal mouse.

A conditional knockout allows gene deletion in a tissue or time specific manner. This is done by introducing short sequences called loxP sites around the gene. These sequences will be introduced into the germ-line via the same mechanism as a knock-out. This germ-line can then be crossed to another germline containing Cre-recombinase which is a viral enzyme that can recognize these sequences, recombines them and deletes the gene flanked by these sites. Because the desired type of DNA recombination is a rare event in the case of most cells and most constructs, the foreign sequence chosen for insertion usually includes a reporter. This enables easy selection of cells or individuals in which knockout was successful. Sometimes the DNA construct inserts into a chromosome without the desired homologous

recombination with the target gene. To eliminate such cells, the DNA construct often contains a second region of DNA that allows such cells to be identified and discarded.

In diploid organisms, which contain two alleles for most genes, and may as well contain several related genes that collaborate in the same role, additional rounds of transformation and selection are performed until every targeted gene is knocked out. Selective breeding may be required to produce homozygous knockout animals.

Gene knockin is similar to gene knockout, but it replaces a gene with another instead of deleting it.

Use

Knockouts are primarily used to understand the role of a specific gene or DNA region by comparing the knockout organism to a wildtype with a similar genetic background.

Knockouts organisms are also used as screening tools in the development of drugs, to target specific biological processes or deficiencies by using a specific knockout, or to understand the mechanism of action of a drug by using a library of knockout organisms spanning the entire genome, such as in *Saccharomyces cerevisiae*

Gene knockout

A gene knockout is a genetically engineered organism that carries one or more genes in its chromosomes that have been made inoperative (have been "knocked out" of the organism). This is done for research purposes. Also known as knockout organisms or simply knockouts, they are used in learning about a gene that has been sequenced, but which has an unknown or incompletely known function. Researchers draw inferences from the difference between the knockout organism and normal individuals. The term also refers to the process of creating such an organism, as in "knocking out" a gene. Knockout is accomplished through a combination of techniques, beginning in the test tube with a plasmid, a bacterial artificial chromosome or other DNA construct, and proceeding to cell culture. Individual cells are genetically transformed with the construct and--for knockouts in multi-cellular organisms--ultimately fused with a stem cell from a nascent embryo. The construct is engineered to recombine with the target gene, which is accomplished by incorporating sequences from the gene itself into the construct. Recombination then occurs in the region of that sequence within the gene, resulting in the insertion of a foreign sequence to disrupt the gene. With its sequence interrupted, the altered gene in most cases will be translated into a nonfunctional protein, if it is translated at all. A conditional knockout allows gene deletion in a tissue specific manner. Because recombination is a rare event in the case of most cells and most constructs, the foreign sequence chosen for insertion usually is a reporter. This enables easy selection of cells or individuals in which knockout was successful. In diploid organisms, which contain two alleles for most genes, and may as well contain several related genes that collaborate in the same role, additional rounds of transformation and selection are performed until every targeted gene is knocked out. Knock-in is similar to knock-out, but instead it replaces a gene with another instead of deleting it.

Knockout mouse A knockout mouse is a genetically engineered mouse one or more of whose genes have been made inoperable through a gene knockout. Knockout is a route to learning about a gene that has been sequenced but has an unknown or incompletely known function. Mice are the laboratory animal species most closely related to humans in which the knockout technique can be easily performed, so they are a favourite subject for knockout experiments, especially with regard to genetic questions that relate to human physiology. (Gene knockout in rats is much harder and has only been possible since 2003.) Use Knocking out the activity of a gene provides information about what that gene normally does. Humans share many genes with mice. Consequently, observing the characteristics of knockout mice gives researchers information that can be used to better understand how a similar gene may cause or contribute to disease in

humans. Examples of research in which knockout mice have been useful include studying and modelling different kinds of cancer, obesity, heart disease, diabetes, arthritis, substance abuse, anxiety, aging and Parkinson disease. Knockout mice also offer a biological context in which drugs and other therapies can be developed and tested. Many of these mouse models are named after the gene that has been inactivated. For example, the p53 knockout mouse is named after the p53 gene which codes for a protein that normally suppresses the growth of tumours by arresting cell division. Humans born with mutations that inactivate the p53 gene suffer from Li-Fraumeni syndrome, a condition that dramatically increases the risk of developing bone cancers, breast cancer and blood cancers at an early age. Other mouse models are named, often with creative flair, according to their physical characteristics or behaviours. For example, "Methuselah" is a knockout mouse model noted for longevity, while "Frantic" is a model useful for studying anxiety disorders.

Procedure There are several variations to the procedure of producing knockout mice; the following is a typical example.

1. The gene to be knocked out is isolated from a mouse gene library. Then a new DNA sequence is engineered which is very similar to the original gene and its immediate neighbor sequence, except that it is changed sufficiently to make it inoperable. Usually, the new sequence is also given a marker gene, a gene that normal mice don't have and that transfers resistance to a certain antibiotic or a selectable marker.
2. From a mouse morula (a very young embryo consisting of a ball of undifferentiated cells), stem cells are isolated; these can be grown in vitro. For this example, we will take a stem cell from a white mouse.
3. The stem cells from step 2 are combined with the new sequence from step 1. This is done via electroporation (using electricity to transfer the DNA across the cell membrane). Some of the electroporated stem cells will incorporate the new sequence into their chromosomes in place of the old gene; this is called homologous recombination. The reason for this process is that the new and the old sequence are very similar. Using the antibiotic from step 1, those stem cells that actually did incorporate the new sequence can be quickly isolated from those that did not.
4. The stem cells from step 3 are inserted into mouse blastocyst cells. For this example, we use blastocysts from a grey mouse. These blastocysts are then implanted into the uterus of female mice, to complete the pregnancy. The blastocysts contain two types of stem cells: the original ones (grey mouse), and the newly engineered ones (white mouse). The newborn mice will therefore be chimeras: parts of their bodies result from the original stem cells, other parts result from the engineered stem cells. Their furs will show patches of white and grey.
5. Newborn mice are only useful if the newly engineered sequence was incorporated into the germ cells (egg or sperm cells). So we cross these new mice with others

and watch for offspring that are all white. These are then further inbred to produce mice that carry no functional copy of the original gene.

Limitations

While knockout mice technology represents a valuable research tool, some important limitations exist. About 15 percent of gene knockouts are developmentally lethal, which means that the genetically altered embryos cannot grow into adult mice. The lack of adult mice limits studies to embryonic development and often makes it more difficult to determine a gene's function in relation to human health. In some instances, the gene may serve a different function in adults than in developing embryos. Knocking out a gene also may fail to produce an observable change in a mouse or may even produce different characteristics from those observed in humans in which the same gene is inactivated. For example, mutations in the p53 gene are associated with more than half of human cancers and often lead to tumours in a particular set of tissues. However, when the p53 gene is knocked out in mice, the animals develop tumours in a different array of tissues. There is variability in the whole procedure depending largely on the strain from which the stem cells have been derived. Generally cells derived from strain 129 are used. This specific strain is not suitable for many experiments (e.g., behavioural), so it is very common to backcross the offspring to other strains. Some genomic loci have been proven very difficult to knock out. Reasons might be the presence of repetitive sequences, extensive DNA methylation, or heterochromatin.

Antisense RNA technology

Antisense RNA (asRNA) is a single-stranded RNA that is complementary to a messenger RNA (mRNA) strand transcribed within a cell. Some authors have used the term micRNA (mRNA-interfering complementary RNA) to refer to these RNAs but it is not widely used.^[1]

Antisense RNA may be introduced into a cell to inhibit translation of a complementary mRNA by base pairing to it and physically obstructing the translation machinery.^[2] This effect is therefore stoichiometric. An example of naturally occurring mRNA antisense mechanism is the *hok/sok* system of the *E. coli* R1 plasmid. Antisense RNA has long been thought of as a promising technique for disease therapy; the first antisense therapeutic to reach the market is the drug fomivirsen, approved in 1998. Mipomersen was approved in the United States in 2013. One commentator has characterized antisense RNA as one of "dozens of technologies that are gorgeous in concept, but exasperating in [commercialization]".^[3] Generally, antisense RNA still lack effective design, biological activity, and efficient route of administration.^[4]

The effects of antisense RNA are related with the effects of RNA interference (RNAi). The RNAi process, found only in eukaryotes, is initiated by double-stranded RNA fragments, which may be created by the expression of an anti-sense RNA followed by the base-pairing of the anti-sense strand to the target transcript.^[5] Double-stranded RNA may be created by other mechanisms (including secondary RNA structure). The double-stranded RNA is cleaved into small fragments by DICER, and then a single strand of the fragment is incorporated into the RNA-induced silencing complex (RISC) so that the RISC may bind to and degrade the complementary mRNA target.^[6] Some genetically engineered transgenic plants that express antisense RNA do activate the RNAi pathway.^[7] This processes resulted in differing magnitudes of gene silencing induced by the expression of antisense RNA. Well-known examples include the FlavrSavr tomato and two cultivars of ringspot-resistant papaya.^{[8][9]}

Transcription of longer *cis*-antisense transcripts is a common phenomenon in the mammalian transcriptome.^[10] Although the function of some cases have been described, such as the *Zeb2/Sip1* antisense RNA, no general function has been elucidated. In the case of *Zeb2/Sip1*,^[11] the antisense noncoding RNA is opposite the 5' splice site of an intron in the 5'UTR of the *Zeb2* mRNA. Expression of the antisense ncRNA prevents splicing of an intron that contains a ribosome entry site necessary for efficient expression of the *Zeb2* protein. Transcription of long antisense ncRNAs is often concordant with the associated protein-

coding gene,^[12] but more detailed studies have revealed that the relative expression patterns of the mRNA and antisense ncRNA are complex

Organization of prokaryotic and eukaryotic genomes

Prokaryotic

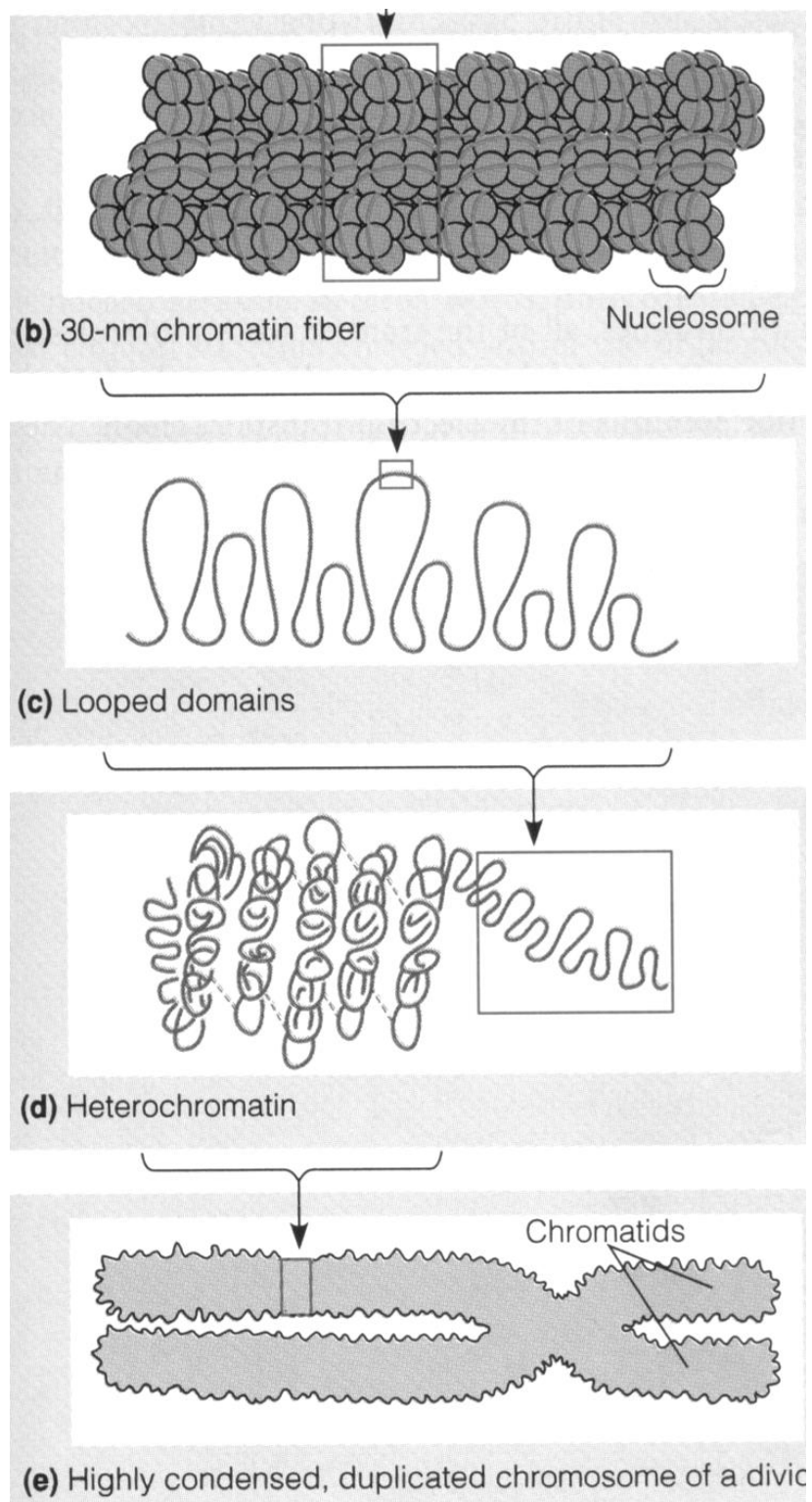
- Usually circular
- Smaller
- Found in the nucleoid region
- Less elaborately structured and folded

Eukaryotic

- Complexed with a large amount of protein to form chromatin
- Highly extended and tangled during interphase
- Found in the nucleus

The current model for progressive levels of DNA packing:

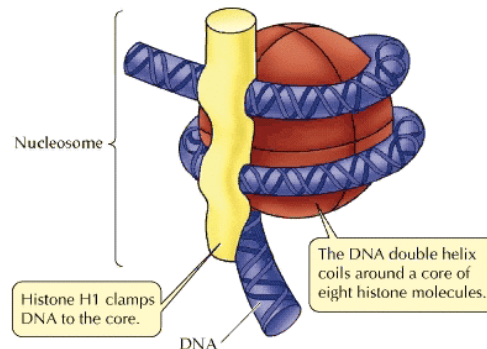
- Nucleosome → basic unit of DNA packing formed from DNA wound around a protein core that consists of 2 copies each of the 4 types of histone (H2A, H2B, H3, H4)]
- A 5th histone (H1) attaches near the bead when the chromatin undergoes the next level of packing
- 30 nm chromatin fiber → next level of packing; coil with 6 nucleosomes per turn
- the 30 nm chromatin forms looped domains, which are attached to a nonhistone protein scaffold (contains 20,000 – 100,000 base pairs)
- Looped domains attach to the inside of the nuclear envelope the 30 nm chromatin forms looped domains, which are attached to a nonhistone protein scaffold (contains 20,000 – 100,000 base pairs)



Histones influence folding in eukaryotic DNA.

- Histones → small proteins rich in basic amino acids that bind to DNA, forming chromatin

Contain a high proportion of positively charged amino acids which bind tightly to the negatively charged DNA



Heterochromatin

- Chromatin that remains highly condensed during interphase and is NOT actively transcribed

Euchromatin

- Chromatin that is less condensed during interphase and IS actively transcribed
- Becomes highly condensed during mitosis

Satellite DNA

→ highly repetitive DNA consisting of short unusual nucleotide sequences that are tandemly repeated 1000's of times

- It is found at the tips of chromosomes and the centromere

Its function is not known, perhaps it plays a structural role during chromosome replication and separation

Table 19.1 Types of Repetitive DNA**Tandemly Repetitive DNA (Satellite DNA)**

Repeated units at a site are usually identical

Proportion of mammalian DNA:	10–15%
Length of each repeated unit:	1–10 base pairs
Total length of repetitive DNA per site, in base pairs:	
Regular satellite DNA	100,000–10 million
Minisatellite DNA	100–100,000
Microsatellite DNA	10–100

Interspersed Repetitive DNA

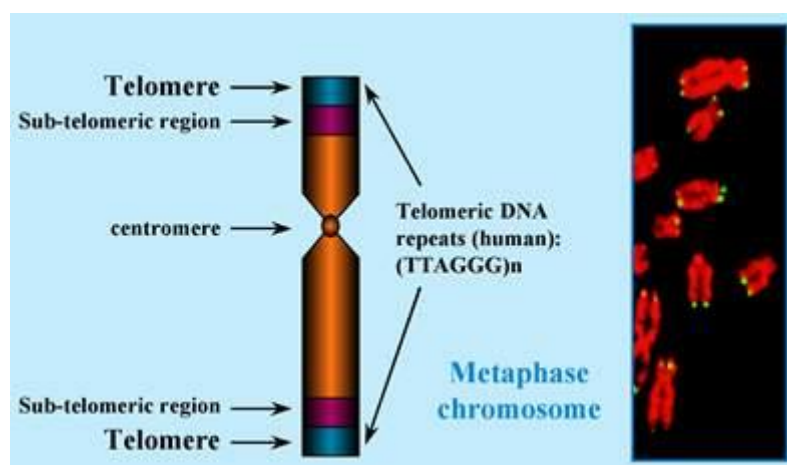
“Copies” are very similar but not identical

Proportion of mammalian DNA:	25–40%
Length of each repeated unit:	100–10,000 base pairs
Number of repetitions per genome:	10–1 million

Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Telomere→ series of short tandem repeats at the ends of eukaryotic chromosomes; prevents chromosomes from shortening with each replication cycle

Telomerase→ enzyme that periodically restores this repetitive sequence to the ends of DNA molecules



Genome Packaging in Prokaryotes: the Circular Chromosome of *E. coli*

E. coli: A Model Prokaryote

Much of what is known about prokaryotic chromosome structure was derived from studies of *Escherichia coli*, a bacterium that lives in the human colon and is commonly used in laboratory cloning experiments. In the 1950s and 1960s, this bacterium became the model organism of choice for prokaryotic research when a group of scientists used phase-contrast microscopy and autoradiography to show that the essential genes of *E. coli* are encoded on a

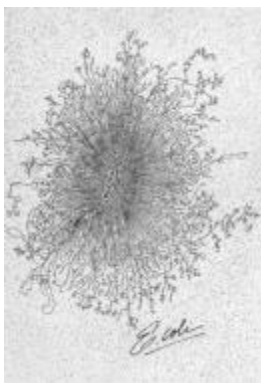
single circular chromosome packaged within the cell nucleoid (Mason & Powelson, 1956; Cairns, 1963).

Prokaryotic cells do not contain nuclei or other membrane-bound organelles. In fact, the word "prokaryote" literally means "before the nucleus." The nucleoid is simply the area of a prokaryotic cell in which the chromosomal DNA is located. This arrangement is not as simple as it sounds, however, especially considering that the *E. coli* chromosome is several orders of magnitude larger than the cell itself. So, if bacterial chromosomes are so huge, how can they fit comfortably inside a cell—much less in one small corner of the cell?

DNA Supercoiling

The answer to this question lies in DNA packaging. Whereas eukaryotes wrap their DNA around proteins called histones to help package the DNA into smaller spaces, most prokaryotes do not have histones (with the exception of those species in the domain Archaea). Thus, one way prokaryotes compress their DNA into smaller spaces is through supercoiling (Figure 1). Imagine twisting a rubber band so that it forms tiny coils. Now twist it even further, so that the original coils fold over one another and form a condensed ball. When this type of twisting happens to a bacterial genome, it is known as supercoiling. Genomes can be negatively supercoiled, meaning that the DNA is twisted in the opposite direction of the double helix, or positively supercoiled, meaning that the DNA is twisted in the same direction as the double helix. Most bacterial genomes are negatively supercoiled during normal growth.

Proteins Involved in Supercoiling



During the 1980s and 1990s, researchers discovered that multiple proteins act together to fold and condense prokaryotic DNA. In particular, one protein called HU, which is the most abundant protein in the nucleoid, works with an enzyme called topoisomerase I to bind DNA and introduce sharp bends in the chromosome, generating the tension necessary for negative supercoiling. Recent studies have also shown that other proteins, including integration host

factor (IHF), can bind to specific sequences within the genome and introduce additional bends (Rice *et al.*, 1996). The folded DNA is then organized into a variety of conformations (Sinden & Pettijohn, 1981) that are supercoiled and wound around tetramers of the HU protein, much like eukaryotic chromosomes are wrapped around histones (Murphy & Zimmerman, 1997).

Once the prokaryotic genome has been condensed, DNA topoisomerase I, DNA gyrase, and other proteins help maintain the supercoils. One of these maintenance proteins, H-NS, plays an active role in transcription by modulating the expression of the genes involved in the response to environmental stimuli. Another maintenance protein, factor for inversion stimulation (FIS), is abundant during exponential growth and regulates the expression of more than 231 genes, including DNA topoisomerase I (Bradley *et al.*, 2007).

Accessing Supercoiled Genes

Supercoiling explains how chromosomes fit into a small corner of the cell, but how do the proteins involved in replication and transcription access the thousands of genes in prokaryotic chromosomes when everything is packaged together so tightly? It has been determined that prokaryotic DNA replication occurs at a rate of 1,000 nucleotides per second, and prokaryotic transcription occurs at a rate of about 40 nucleotides per second (Lewin, 2007), so bacteria must have highly efficient methods of accessing their DNA strands. But how?

Researchers have noted that the nucleoid usually appears as an irregularly shaped mass within the prokaryotic cell, but it becomes spherical when the cell is treated with chemicals to inhibit transcription or translation. Moreover, during transcription, small regions of the chromosome can be seen to project from the nucleoid into the cytoplasm (i.e., the interior of the cell), where they unwind and associate with ribosomes, thus allowing easy access by various transcriptional proteins (Dürrenberger *et al.*, 1988). These projections are thought to explain the mysterious shape of nucleoids during active growth. When transcription is inhibited, however, the projections retreat into the nucleoid, forming the aforementioned spherical shape.

Because there is no nuclear membrane to separate prokaryotic DNA from the ribosomes within the cytoplasm, transcription and translation occur simultaneously in these organisms. This is strikingly different from eukaryotic chromosomes, which are confined to the membrane-bound nucleus during most of the cell cycle. In eukaryotes, transcription must be completed in the nucleus before the newly synthesized mRNA molecules can be transported to the cytoplasm to undergo translation into proteins.

Variations in Prokaryotic Genome Structure

Recently, it has become apparent that one size does not fit all when it comes to prokaryotic chromosome structure. While most prokaryotes, like *E. coli*, contain a single circular DNA molecule that makes up their entire genome, recent studies have indicated that some prokaryotes contain as many as four linear or circular chromosomes. For example, *Vibrio cholerae*, the bacteria that causes cholera, contains two circular chromosomes. One of these chromosomes contains the genes involved in metabolism and virulence, while the other contains the remaining essential genes (Trucksis *et al.*, 1998). An even more extreme example is provided by *Borrelia burgdorferi*, the bacterium that causes Lyme disease. This organism is transmitted through the bite of deer ticks (Figure 2), and it contains up to 11 copies of a single linear chromosome (Ferdows & Barbour, 1989). Unlike *E. coli*, *Borrelia* cannot supercoil its linear chromosomes into a tight ball within the nucleoid; rather, these strands are diffused throughout the cell.

Other organisms, such as *Bacillus subtilis*, form nucleoids that closely resemble those of *E. coli*, but they use different architectural proteins to do so. Furthermore, the DNA molecules of Archaea, a taxonomic domain composed of single-celled, nonbacterial prokaryotes that share many similarities with eukaryotes, can be negatively supercoiled, positively supercoiled, or not supercoiled at all. It is important to note that archaeans are the only group of prokaryotes that use eukaryote-like histones, rather than the architectural proteins described above, to condense their DNA molecules (Sandman *et al.*, 1990). The acquisition of histones by archaeans is thought to have paved the way for the evolution of larger and more complex eukaryotic cells.

Other DNA Differences Between Prokaryotes and Eukaryotes

Most prokaryotes reproduce asexually and are haploid, meaning that only a single copy of each gene is present. This makes it relatively easy to generate mutations in the lab and study the resulting phenotypes. By contrast, eukaryotes that reproduce sexually generally contain multiple chromosomes and are said to be diploid, because two copies of each gene exist—with one copy coming from each of an organism's parents.

Yet another difference between prokaryotes and eukaryotes is that prokaryotic cells often contain one or more plasmids (i.e., extrachromosomal DNA molecules that are either linear or circular). These pieces of DNA differ from chromosomes in that they are typically smaller and encode nonessential genes, such as those that aid growth in specific conditions or encode antibiotic resistance. *Borrelia*, for instance, contains more than 20 circular and linear

plasmids that encode genes responsible for infecting ticks and humans (Fraser *et al.*, 1997). Plasmids are often much smaller than chromosomes (i.e., less than 1,500 kilobases), and they replicate independently of the rest of the genome. However, some plasmids are capable of integrating into chromosomes or moving from cell to cell.

Perhaps due to the space constraints of packing so many essential genes onto a single chromosome, prokaryotes can be highly efficient in terms of genomic organization. Very little space is left between prokaryotic genes. As a result, noncoding sequences account for an average of 12% of the prokaryotic genome, as opposed to upwards of 98% of the genetic material in eukaryotes (Ahnert *et al.*, 2008). Furthermore, unlike eukaryotic chromosomes, most prokaryotic genomes are organized into polycistronic operons, or clusters of more than one coding region attached to a single promoter, separated by only a few base pairs. The proteins encoded by each operon often collaborate on a single task, such as the metabolism of a sugar into by-products that can be used for energy (Figure 3).

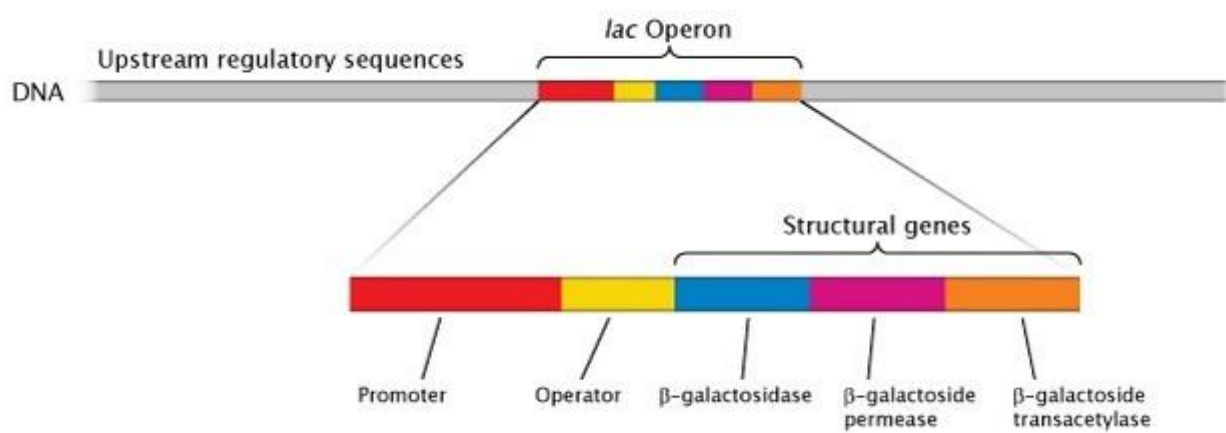


Figure 3: The prokaryotic *lac* operon.

Three structural genes code for proteins involved in lactose import and metabolism in bacteria. The genes are organized together in a cluster called the *lac* operon.

© 2013 Nature Education All rights reserved. 

Figure Detail

The organization of prokaryotic DNA therefore differs from that of eukaryotes in several important ways. The most notable difference is the condensation process that prokaryotic DNA molecules undergo in order to fit inside relatively small cells. Other differences, while not as dramatic, are summarized in Table 1.

Table 1: Prokaryotic versus Eukaryotic Chromosomes

Prokaryotic Chromosomes	Eukaryotic Chromosomes
<p>Many prokaryotes contain a single circular chromosome.</p> <p>Prokaryotic chromosomes are condensed in the nucleoid via DNA supercoiling and the binding of various architectural proteins.</p> <p>Because prokaryotic DNA can interact with the cytoplasm, transcription and translation occur simultaneously.</p> <p>Most prokaryotes contain only one copy of each gene (i.e., they are haploid).</p> <p>Nonessential prokaryotic genes are commonly encoded on extrachromosomal plasmids.</p> <p>Prokaryotic genomes are efficient and compact, containing little repetitive DNA.</p>	<p>Eukaryotes contain multiple linear chromosomes.</p> <p>Eukaryotic chromosomes are condensed in a membrane-bound nucleus via histones.</p> <p>In eukaryotes, transcription occurs in the nucleus, and translation occurs in the cytoplasm.</p> <p>Most eukaryotes contain two copies of each gene (i.e., they are diploid).</p> <p>Some eukaryotic genomes are organized into operons, but most are not.</p> <p>Extrachromosomal plasmids are not commonly present in eukaryotes.</p> <p>Eukaryotes contain large amounts of noncoding and repetitive DNA.</p>

Genome Projects

Genome projects are scientific endeavours that ultimately aim to determine the complete genome sequence of an organism (be it an animal, a plant, a fungus, a bacterium, an archaean, a protist or a virus) and to annotate protein-coding genes and other important genome-encoded features. The genome sequence of an organism includes the collective DNA sequences of each chromosome in the organism. For a bacterium containing a single chromosome, a genome project will aim to map the sequence of that chromosome. For the human species, whose genome includes 22 pairs of autosomes and 2 sex chromosomes, a complete genome sequence will involve 46 separate chromosome sequences.

What is genome science? (or genomics?) : It is the study of the structure, content, and evolution of genome.

Genome Projects

- Scientific endeavors that ultimately aim to determine the complete genome sequence of an organism.

- aim to increase understanding of genetics

- increase dimensionality of where and when

- Ultimate goal is establish an integrated web-based database and research interface

"Real" goals of genome projects

- To generate and order genomic and expressed gene sequences

- To identify and annotate genes of a genome

- To characterize DNA sequence diversity

Functional Genomics

- finding the meaning of the genome

- What are the biochemical, cellular, and/or physiological properties of each and every gene product of a genome?

- genome-scale gene expression profile

- genome-scale protein expression profile

- structural genomics, pharmacogenics

Comparative Genomics

Physical attributes of the DNA with locus information that may or may not include phenotypes.

Unit of genetic distance •centiMorgan

Physical Map

- An assembly of contiguous stretches of chromosomal DNA, aka "contigs".
(expressed in Kb)
- Work complementarily with genetic maps

Two general strategies to assemble contigs

- 1.Alignment of randomly isolated clones based on RFLP marker locations
- 2.Chromosomal walking

Cytological Maps

Cytological maps are the banding patterns observed through a microscope on stained chromosome spreads. Can be aligned with a physical map using insitu hybridization of cloned DNA fragments to the chromosomes a technique called FISH

Syntenry

A phenomenon in which local gene order along a chromosome tends to be conserved among different related species. Can be used to identify orthologs.

Chromosomal painting

A powerful method for determining syntenry

Orthologs

diverged by evolutionary split or speciation. We can use orthologs as anchoring landmarks for assembling a genome using syntenic segments from other species.

paralogs

diverged by gene duplication

Gene genealogy

An ultimate test for orthologous vs. paralogous relationships between homologs.

Assembling a genetic map

Smaller distance is much better than large distance because it can be prone to mutations/crossover.

Linkage group

Essentially a chromosome, which is ~100cM in animals; one crossover per chromosome per generation.

Genetic Map

The relative order of genetic markers in linkage groups in which the distance between markers is expressed as units of recombination.

Genetic markers

physical attributes of the DNA with locus information that may or may not include phenotypes. (With phenotypes: drosophila white mutation; without phenotypes: RFLP microsatellite markers)

Plant Genome projects

Since the publication in 2000 of the model *Arabidopsis thaliana* genome in the journal *Nature*, the number of genomes has steadily increased, peaking in 2012 with 13 publications (Fig. 1A). At this current trajectory there should be hundreds of plant genome publications over the next several years. Genome papers have been quite formulaic with a description of the assembly, gene numbers, repeats, WGDs, over and under-represented gene families, and finally, some aspect of novel biology, usually with a focus on transcription factors. Genomes have been published in 12 different journals with 38 of the 55 (69%) published genomes appearing in Nature journals (*Nature*, *Nature Genetics*, *Nature Biotech*, and *Nature Communications*); *Science* is second with six published genomes. As we see from the most recent publication of the *Capsella rubella* genome paper, the genome paper is shifting from a formulaic approach to a focus on how the genome elucidates novel biological aspects, such as the evolution of selfing to an outcrossing mating system (Slotte et al., 2013). The trend toward biology is quite positive and necessitated by demands for publication in high impact journals. However, the plant community is just at the beginning of exploring the diversity of plant genomes, and the rigor of the genome paper model with the associated in-depth exploration of genome features provides an essential foundation for the plant research community.

Table 1.

[View Full Table](#) | [Close Full View](#)

Published plant genomes.†

	Scientific name	Common name	Year	Type	Divisio
1	<i>Arabidopsis thaliana</i>	arabidopsis	2000	model	
2	<i>Oryza sativa</i>	rice	2002	crop	
3	<i>Oryza sativa</i>	rice	2002	crop	
4	<i>Oryza sativa</i>	rice	2005	crop	
5	<i>Populus trichocarpa</i>	black cottonwood	2006	crop	
6	<i>Vitis vinifera</i>	grape	2007	crop	
7	<i>Physcomitrella patens</i>	moss	2008	model	
8	<i>Vitis vinifera</i>	grape	2007	crop	
9	<i>Carica papaya</i>	papaya	2008	crop	
10	<i>Lotus japonicus</i>	lotus	2008	model	
11	<i>Sorghum bicolor</i>	sorghum	2009	crop	
12	<i>Cucumis sativus</i>	cucumber	2009	crop	
13	<i>Zea mays</i>	maize	2009	crop	
14	<i>Glycine max</i>	soybean	2010	crop	
15	<i>Brachypodium distachyon</i>	brachypodium	2010	model	
16	<i>Ricinus communis</i>	castor bean	2010	crop	
17	<i>Malus x domestica</i>	apple	2010	crop	
18	<i>Jatropha curcas</i>	jatropha	2010	crop	
19	<i>Theobroma cacao</i>	cocoa	2011	crop	
20	<i>Fragaria vesca</i>	strawberry	2011	crop	

One of the forces driving the rapid increase in fully sequenced plant genomes is the exponential decrease in cost and speed of genome sequencing fueled by high throughput DNA sequencing (Schatz et al., 2012). More than half of the published genomes have been sequenced entirely or partly using Sanger technology (Table 1), which provides long high quality ~1000 base pair (bp) reads. Sanger sequencing requires a cloning step and is time consuming with an expensive price tag, although the final result is usually high quality depending on the genome. When 454 came onto the scene in the early 2000s the cost of sequencing dropped an order of magnitude (US\$200K vs. US\$2 M) encouraging the emergence of consortia and funding for the sequencing of new genomes. Grape was the first genome published in 2007 using a combination of 454 and Sanger, and now there are at least 18 genomes that have used varying amounts of 454 sequence. Illumina and SOLiD sequencing changed the paradigm yet again providing very short reads (35–150 bp) at yet another order of magnitude lower cost than 454. Only two genome projects have used SOLiD for genome sequencing (strawberry and tomato); however, Illumina has played an exclusive

role in 12 genomes, and was used in combination with other technologies in another 17 genomes. Third generation sequencing technologies such as Pacific Bioscience (PacBio) promise long (>5 kb) single molecule reads that would greatly improve assembly of repeat rich plant genomes. PacBio long reads show great promise in resolving regions that the other sequencing technologies have problems with (skewed GC, homopolymers), but throughput and accuracy are two issues that still require attention. However, new sequencing technologies are only part of the future of plant genomes since tried and true methods, such as BACs (bacterial artificial chromosomes), are finding a place in hybrid sequencing approaches such as in the highly heterozygous pear genome (Wu et al., 2013).

Most of the plants chosen to be sequenced to date fit specific criteria such as size of research community, model organisms or economically important, small genome size, ploidy (diploid), availability of inbred lines (low heterozygosity), access to genetic and physical maps, expressed sequence tags (EST)/transcriptome and other genomic tools. Seventy-three percent (40) of the plant genome publications have been on crop species and some of these crop species double as model systems while several were sequenced purely for research such as *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Brachypodium distachyon*, *Physcomitrella patens* (moss), and *Selaginella moellendorffii* (spikemoss). Most (94%) genomes sequenced to date are Angiosperms, of which 36 are dicots and 16 are monocots, while only one gymnosperm (spruce), one bryophyte (moss), and one lycophyta (club-moss) have been sequenced (Table 1). Much of the early decisions about which genomes to sequence were driven by the Department of Energy Joint Genome Institute (JGI), which resulted in the publication and public availability (phytozome) of eleven of the highest quality plant genomes. The Beijing Genome Institute (BGI) has contributed consistently over the years starting with the rice genome, then ten additional genomes primarily based on Illumina technology, and now they have announced a large-scale plant genome sequencing project. However, a “1000 plant genome project” analogous to that in other communities has yet to emerge.

Plant Genomes both Large and Small

Plant genome sizes span several orders of magnitude from the carnivorous corkscrew plant (*Genlisea aurea*) at 63 megabases (Mb) to the rare Japanese *Paris japonica* at 148,000 Mb (Bennett and Leitch, 2011). The smallest published genome is the carnivorous bladderwort (*Utricularia gibba*) at 82 Mb, while the largest, the Norway Spruce (*Picea abies*), stands by itself at 19,600 Mb, compared to the second largest of maize at 2300 Mb and the overall median of 480 Mb (Table 1, Fig. 1B). Access to high quality reference genomes confirmed that long terminal repeats (LTRs) retrotransposons are a primarily driver of the dramatic size range in plants (El Baidouri and Panaud 2013). For the large barley genome (5100 Mb), where retrotransposons are abundant and more recently active, a powerful genomics resource was generated through an alternative “gene-ome” approach by anchoring a high quality genespace assembly on a deep physical map merged with high-density genetic maps (International Barley Genome Sequencing Consortium, 2012). In contrast, large gymnosperm genomes have highly diverged ancient repeats, which could make assembling these genomes tractable with current sequencing and assembly technologies (Kovach et al., 2010). The smallest reported conifer genome is the same size as maize and the median genome size is 9700 Mb, which is why a large push to sequence gymnosperms may have to wait for the next wave of sequencing technologies with increased read length and decreased price. As the community moves forward to choose the next round of genomes to sequence, the Kew Genome Size database will continue to provide a rich resource of non-model/non-crop species to investigate (Bennett and Leitch, 2011).

One measure of genome assembly quality is the contiguity or the length of contigs and scaffolds at which 50% of the assembly can be found; this is commonly referred to as N50. Sorghum, *Brachypodium distachyon*, soybean, and foxtail millet have the top four scaffold contiguities with 62.4, 59.3, 47.8, and 47.3 Mb respectively and all four were sequenced using Sanger as part of the JGI pipeline (Table 1). However, the genome with the ninth largest scaffold N50 is the tomato genome at 16 Mb, which was predominantly assembled using 454. Each scaffold is comprised of thousands of contigs and contig length generally drives the completeness and quality of the gene predictions. Not surprising, the 11 JGI assemblies based on Sanger have the top contig N50 ranging from 347 to 119 kilobases (kb), while the median contig N50 for all assemblies is 25.6 kb. Illumina based assemblies, primarily from BGI, have a similar median length (25.9 kb), which reflects their comprehensive strategy that makes use of different sized sequencing libraries. Another measure of a genome assembly is the amount of the genome captured in the assembly. Of the

published genomes, the median genome assembly captured 85% of the predicted genome size, which is usually estimated by flow cytometry or more recently by k-mer depth analysis. The remaining fraction of the genome not assembled generally represents the highly repetitive portion of the genome such as high copy number ribosomal repeats, centromeres, telomeres, and transposable elements. Therefore an average plant genome assembly captures 85% of the genome space in thousands of contigs with an N50 of 20 kb and tens of scaffolds with an N50 of 1 Mb.

Annotation of any genome, but particularly plant genomes, is difficult especially as the definition of what constitutes a gene continues to evolve. Many parts of the genome are ‘expressed’ in that RNAs are formed, but do not correspond to traditional genes in that they are not translated to a protein. However, most annotated plant genomes have between 20,000 and 94,000 genes with a median predicted gene count of 32,605 (Table 1, Fig. 1C). Differences between genomes most likely lies in the tools used for annotation and how relaxed the annotators were in calling genes as well as lineage-specific genes and gene family expansions. Genomes produced by next generation sequencing typically have smaller contig and scaffold sizes that complicate annotation as genes may not exist on single contigs but may be broken across contigs, thus inflating the number of annotated genes (e.g., pigeon pea, Varshney et al., 2012). Further complicating annotation is that there are many expressed non-coding RNAs that are functionally important (Eddy, 2001), but not considered genes in a traditional sense. Small RNA precursors are often not included in a genome annotation, but are important for plant development and silencing of TEs (Arikait et al., 2013). Small RNAs and other non-coding RNAs are often annotated and curated separately from genome annotations in small, boutique databases. Long-term, however, one goal should be to combine these various sources of information into a single database/annotation making it easier for the biologist to pull together relevant information needed for forming hypotheses.

Plant genomes are packed, and often obese, with transposable elements (TEs) (Bennetzen 2000), which contain protein-coding sequences that are often annotated as genes. In rice, for instance, it was estimated that only 40,000 of the more than 55,000 annotated genes are really genes and that the other 10,000 to 15,000 are TEs—usually low copy TEs as high copy elements are relatively easy to find (Bennetzen et al., 2004). TEs include various families that move via copy-and-paste (class I) and cut-and-paste (class II) mechanisms. Copy-and-paste TEs can dramatically increase the size of a genome such as occurred in a relative of rice with a genome nearly two-fold larger than rice (Piegu et al., 2006). Transposon biology is an intriguing area of research and relies on relatively complete

genomes so that TEs are captured in sequence contigs and can be accurately annotated. Schemes for classification of TEs have been agreed on (Wicker et al., 2007), but annotation of non-LTR TEs is complicated by the lack of structural clues that allow routine ab initio prediction (El Baidouri and Panaud, 2013). Another complication is that in genomes produced by short read DNA sequencing technology, TEs are often missed in the assembly due to their repetitive nature. Genomes sequenced to date range from 3 to 85% repetitive sequence (Table 1; median 43%), with TEs, specifically cut-and-paste TEs (LTRs), comprising the majority of that sequence. Capturing and annotating these genomic components is important as it is becoming increasingly clear that TEs can be domesticated to function in gene regulation and as structural components of the genome.

Making Genomes “Functional”

One of the key take homes from the first 49 sequenced plant species is that we still have a lot to learn about the organization of genomes, function of genes, and how to characterize the non-coding space. Each new genome uncovers novel genes specific to a species, and a vast amount of non-coding space that requires methods for ab initio and functional annotation. One specific challenge is how we will leverage a growing number of high throughput technologies, otherwise referred to as “omics” approaches, to functionally annotate features of the plant genome. In this special issue of *The Plant Genome* we highlight several omics studies that have used high throughput approaches such as gen-omics (SNP detection), epigen-omics (methylation) metagen-omics (plant-fungal interactions), and ion-omics (element profiling) to refine our functional understanding of several key crop genomes (Eichten et al., 2013; Roorkiwal et al., 2013; Ruzicka et al., 2013; Ziegler et al., 2013). As we have seen through the model organism and human ENCODE projects, the layering of omics data exponentially increases the value of a reference genome (Celniker et al., 2009; ENCODE Project Consortium 2012).

While reference genomes provide a starting point, or platform for discovery in a specific species, it only captures a brief moment in the history of that species’ diversity and lacks the information content that would enable activities such as molecular breeding and phylogenetic analyses. Roorkiwal et al. (2013, this issue) describe the development of an Illumina BeadXpress SNP genotyping platform for two important crops in the developing world, pigeon pea and chickpea (Roorkiwal et al., 2013). Both pigeon pea and chickpea have lagged behind other crops in their genetic improvement due to a lack of genome and breeding resources that would enable such applications as marker assisted selection (MAS) and phylogenetic screens to identify genetic novelty in wild species. The development of an

Illumina BeadXpress SNP genotyping platform provides the opportunity to assess larger populations of plants with an adequate density of markers, which is ideal for breeding applications such as MAS and scans of diversity for disease and abiotic traits.

A prominent feature of plant genomes is their epigenetic landscape. The epigenome encompasses DNA methylation, histone modifications and other modifications not directly encoded in the genome. In general, DNA methylation is thought to mark permanent changes in the genome that must exist over the developmental lifetime of the plant, such as silencing transposable elements in embryonic tissue to protect the fidelity of the genome from transposition. Eichten et al. (2013, this issue) address the question of whether DNA methylation also specifies tissue types in maize. Using genome-wide array and sequencing technologies to assess DNA methylation and gene expression in two maize inbreds, B73 and Mo17, across four tissue types (leaf, immature tassel, embryo and endosperm), the authors find that there are more differentially methylated regions (DMRs) between maize inbreds than in the tissues they sampled (Eichten et al., 2013). The DMRs that were identified between tissue types did not correlate with subsequent expression changes suggesting the DMRs were not in fact functional in specifying tissue type. Despite other plants such as tomato that display tissue and developmentally regulated DMRs (Zhong et al., 2013), this may not be a general phenomenon in other species such as maize, which highlights the need to functionally define genomic elements in specific species.

Genetic screens are still the primary tool for functionally defining features of genomes. Mutant screens have been central in elucidating pathways, uncovering novel functionality of known genes, and allowing the discovery of novel non-coding features such as epigenetic regulation and small RNAs. Ziegler et al. (2013, this issue) describe a powerful high throughput mutant screen for elemental differences between field grown soy plants, which could be applied to any plant species with modestly sized seeds like soy (Ziegler et al., 2013). High throughput elemental profiling, or ionomics, is an emerging omics platform that provides a glimpse of a plant–soil environment and how that plant is accessing that environment. Ionomics screens have been powerful at detecting genetic factors controlling ion uptake but also have started to shed light on root architecture and morphology. Therefore, this high throughput screen, which is agnostic to plant species, has the potential to functionally characterize a plant organ, the root, which has traditionally been difficult to define genetically and molecularly in a field environment.

An almost uncharacterized area of plant biology is the complement of organisms that live mutually with plant communities, or the metagenome. In many plants, the acquisition of

inorganic minerals is facilitated by an active network of mycorrhizal associations between soil fungal species and plant roots. However, assessing how these fungal and plant species interact has been hampered by the fact that many fungal species cannot be cultured. The advent of high throughput sequencing has enabled an unprecedented opportunity to identify the genomic changes induced through these communal relationships. Ruzicka et al. (2013, this issue) use high throughput sequencing to characterize the transcriptomes of both the tomato genome and its arbuscular mycorrhizal fungal symbiont in the field (Ruzicka et al., 2013). Instead of culturing the symbiont, a metagenomic sequencing strategy was employed where RNA from a wild-type tomato plant and a mutant for reduced mycorrhizal colonization were sequenced and bioinformatically separated. This metagenomic analysis revealed a suite of genes for transport and cell wall remodeling required for the symbiotic relationship. Metagenomic sequencing will open up the opportunity to explore additional symbiotic relationships and further functionally characterize aspects of the genome that are not innate to the genome sequence.

Future Plant Genomes

The first ~50 plant genomes have provided a glimpse at the gene number, types and numbers of repeats, and how genomes grow and contract. However, we are just at the beginning of defining the functional aspects of plant genomes. To reach the goal of breeding better plants for future food, clothing, and energy, we will need to expand both the species sequenced, the number of species re-sequenced, and the type of omics data layered on genomes. Currently only one gymnosperm has been sequenced and no CAM (Crassulacean acid metabolism) photosynthetic plants have been sequenced. While we have come a long way over the past 13 yr since the publication of the *Arabidopsis* genome, we still have a long way to go before we will be able to engineer the plant of the future.

Insect Genome Projects

The importance of insects

Insects are the largest animal group in the world (75% of all species are insects) and are economically and ecologically extremely important, because most flowering plants depend on insects for their pollination. The honey bees alone, for example, pollinate 15-20 billion dollars worth of crop yearly in the United States.

But insects can also be severe agricultural pests, destroying 30% of our potential annual harvest, and can be vectors (intermediate pathogen carriers) for major diseases such as

malaria, sleeping sickness, Dengue fever, yellow fever, and elephantiasis (for a more complete list of insect-borne diseases see: <http://www.traveldocor.co.uk/insects.htm>). There are an estimated 300-500 million cases of malaria and up to 2.7 million deaths (mainly children) from malaria each year. But also the other diseases are equally serious. Elephantiasis, for example, disables 130 million people worldwide and 1.1 billion people, 17% of the world's population, are at risk of infection.

It will be clear, therefore, that insects can be both very beneficial and harmful (pests) and that a few insect species are hampering the welfare of many hundreds of millions of people especially in the developing countries of the third world. Thus, vast social benefits would be gained, if one selectively could reduce the populations of these pest insects.

Which insects have been sequenced so far?

Although much sequencing efforts have been focused on the twelve *Drosophila* species (1-3), other important insects have also been sequenced during the last eight years (Fig. 1). These insect species are: the malaria mosquito *Anopheles gambiae* (sequenced and published, ref. 4); the silkworm *Bombyx mori* (sequenced and published, refs. 5 and 6); the honey bee *Apis mellifera* (sequenced and published, ref. 7); the red flour beetle *Tribolium castaneum* (sequenced and published, ref. 8); the yellow fever mosquito *Aedes aegypti* (sequenced and published, ref. 9); the mosquito *Culex pipiens* (sequenced, but unpublished); three parasitic wasp species, belonging to the genus *Nasonia* (sequenced, but unpublished); the blood-sucking bug *Rhodnius prolixus* (in progress); the pea aphid *Acyrtosiphon pisum* (in progress); and the body louse *Pediculus humanus* (in progress).

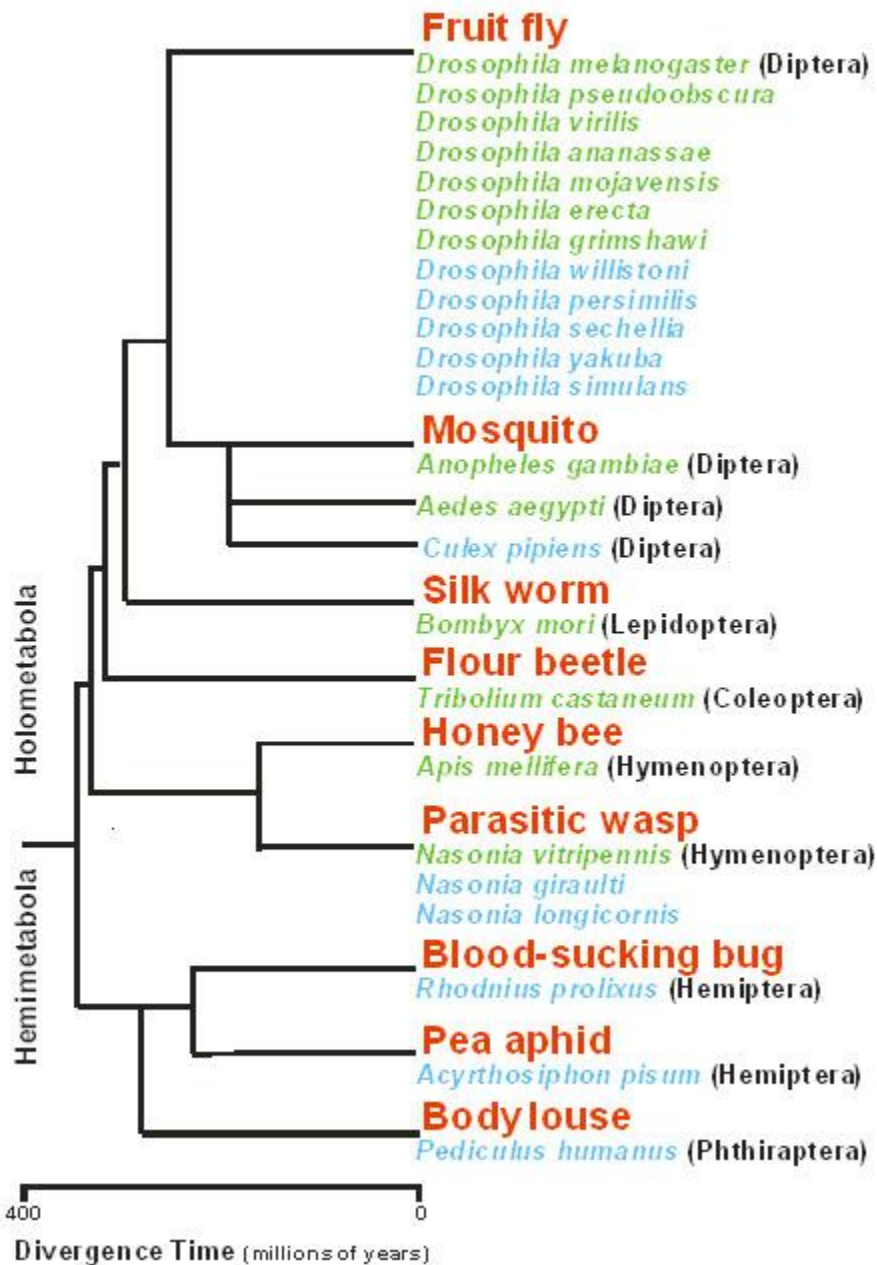


Fig. 1. Phylogenetic relationships of the insects, whose genomes have been sequenced. Green indicates genomes that have been fully sequenced (more than 8x coverage), blue indicates genomes, where the sequencing has not been completed (less than 8x coverage). The sequenced genomes cover about 350 million years of insect evolution.

These twenty-four species represent six different insect orders (Fig. 1) and also the two major evolutionary lineages of insects: Holometabola (insects with a complete metamorphosis) and Hemimetabola (insects having an incomplete metamorphosis). The sequenced genomes from these insects are goldmines, because they contain the information of all proteins and, thereby, of all biochemical and physiological processes that occur in an insect.

Why have these insects been selected for a genome project?

The twenty-four insect species have been selected for a genome project for different reasons. First, some insects are model organisms, such as the fruitfly *D. melanogaster*, which is a well-established model for geneticists and developmental and molecular biologists. The other *Drosophila* species have been sequenced to help with the interpretation of the *D. melanogaster* genome (2, 3).

A second class are the medically important insects that are vectors for serious diseases, such as malaria (the mosquito *A. gambiae*), yellow fever (the mosquito *A. aegypti*), elephantiasis (the mosquito *C. pipiens*), Chagas disease (the blood-sucking bug *R. prolixus*), and typhus (the body louse *P. humanus*).

A third class are the insects that are agriculturally important. To them belong agriculturally beneficial insects, such as the honey bee *A. mellifera*, which is a major pollinator of food plants and producer of honey, and the silkworm *B. mori*, which produces silk. In contrast to them, the red flour beetle *T. castaneum* (which destroys stored grain and many other dried and stored commodities for human consumption) and the pea aphid *A. pisum* (which causes severe damage to green food plants) are serious agricultural pests. The parasitic wasp *N. vitripennis* and the other two *Nasonia* species have been selected, because of their potentials for biological pest control (they lay eggs into a variety of agricultural pest insects).

Human Genome Project (HGP)

The **Human Genome Project (HGP)** was an international scientific research project with the goal of determining the sequence of chemical base pairs which make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and functional standpoint.^[1] It remains the world's largest collaborative biological project.^[2] The project was proposed and funded by the US government; planning started in 1984, got underway in 1990, and was declared complete in 2003. A parallel project was conducted outside of government by the Celera Corporation, or Celera Genomics, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in twenty universities and research centers in the United States, the United Kingdom, Japan, France, Germany, and China.^[3]

The Human Genome Project originally aimed to map the nucleotides contained in a human haploid reference genome (more than three billion). The "genome" of any given individual is unique; mapping "the human genome" involves sequencing multiple variations of each gene.^[4]

History

In May, 1985 Robert Sinsheimer organized a workshop to discuss sequencing the human genome,^[5] but for a number of reasons the NIH was uninterested in pursuing the proposal. The following March, the Santa Fe Workshop was organized by Charles DeLisi and David Smith of the Department of Energy's Office of Health and Environmental Research (OHER).^[6] At the same time Renato Dulbecco proposed whole genome sequencing in an essay in *Science*.^[7] James Watson followed two months later with a workshop held at the Cold Spring Harbor Laboratory.

The fact that the Santa Fe workshop was motivated and supported by a Federal Agency opened a path, albeit a difficult and tortuous one (Cook-Deegan),^[8] for converting the idea into public policy. In a memo to the Assistant Secretary for Energy Research (Alvin Trivelpiece), Charles DeLisi, who was then Director of OHER, outlined a broad plan for the project.^[9] This started a long and complex chain of events which led to approved reprogramming of funds that enabled OHER to launch the Project in 1986, and to recommend the first line item for the HGP, which was in President Regan's 1988 budget submission (Cook-Deegan),^[10] and ultimately approved by the Congress. Of particular importance in Congressional approval was the advocacy of Senator Peter Domenici, whom DeLisi had befriended.^[11] Domenici chaired the Senate Committee on Energy and Natural Resources, as well as the Budget Committee, both of which were key in the DOE budget process. Congress

added a comparable amount to the NIH budget, thereby beginning official funding by both agencies.

Dr. Alvin Trivelpiece sought and obtained the approval of DeLisi's proposal by Deputy Secretary William Flynn Martin. This chart^[12] was used in the spring of 1986 by Trivelpiece, then Director of the Office of Energy Research in the Department of Energy, to brief Martin and Under Secretary Joseph Salgado regarding his intention to reprogram \$4 million to initiate the project with the approval of Secretary Herrington. This reprogramming was followed by a line item budget of \$16 million in the Reagan Administration's 1987 budget submission to Congress.^[13] It subsequently passed both Houses. The Project was planned for 15 years.^[14]

Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.^[15]

In 1990, the two major funding agencies, DOE and NIH, developed a memorandum of understanding in order to coordinate plans and set the clock for the initiation of the Project to 1990.^[16] At that time, David Galas was Director of the renamed "Office of Biological and Environmental Research" in the U.S. Department of Energy's Office of Science and James Watson headed the NIH Genome Program. In 1993, Aristides Patrinos succeeded Galas and Francis Collins succeeded James Watson, assuming the role of overall Project Head as Director of the U.S. National Institutes of Health (NIH) National Center for Human Genome Research (which would later become the National Human Genome Research Institute). A working draft of the genome was announced in 2000 and the papers describing it were published in February 2001. A more complete draft was published in 2003, and genome "finishing" work continued for more than a decade.

The \$3-billion project was formally founded in 1990 by the US Department of Energy and the National Institutes of Health, and was expected to take 15 years.^[17] In addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Australia, China and myriad other spontaneous relationships.^[18]

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by U.S. President Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000).^[19] This first available rough draft assembly of the genome was completed by the Genome Bioinformatics Group at the University of California, Santa Cruz, primarily led by then graduate student Jim Kent. Ongoing sequencing led to the announcement of the essentially

complete genome on April 14, 2003, two years earlier than planned.^{[20][21]} In May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in *Nature*.^[22]

State of completion

The project did not aim to sequence all the DNA found in human cells. It sequenced only "euchromatic" regions of the genome, which make up about 90% of the genome. The other regions, called "heterochromatic" are found in centromeres and telomeres, and were not sequenced under the project.^[23]

The Human Genome Project was declared complete in April 2003. An initial rough draft of the human genome was available in June 2000 and by February 2001 a working draft had been completed and published followed by the final sequencing mapping of the human genome on April 14, 2003. Although this was reported to be 99% of the euchromatic human genome with 99.99% accuracy a major quality assessment of the human genome sequence was published on May 27, 2004 indicating over 92% of sampling exceeded 99.99% accuracy which was within the intended goal.^[24] Further analyses and papers on the HGP continue to occur.^[25]

What are the overall goals of the HGP?

The Human Genome Project has several goals, which include *mapping*, *sequencing*, and *identifying* genes, *storing* and *analyzing* data, and *addressing* the ethical, legal, and social issues (ELSI) that may arise from availability of personal genetic information. *Mapping* is the construction of a series of chromosome descriptions that depict the position and spacing of genes, which are on the DNA of chromosomes. ***The ultimate goal of the Human Genome Project is to decode, letter by letter, the exact sequence of all 3.2 billion nucleotide bases that make up the human genome.*** This means constructing *detailed genetic and physical maps of the human genome*. Besides determining the complete nucleotide sequence of human DNA, this includes locating the genes within the human genome. The HGP agenda also includes analyzing the genomes of several other organisms (including *E. coli*, the fruit fly, and the laboratory mouse) that are used extensively in research laboratories as model systems. Studying the genetic makeup of non-human organisms will help in understanding and deciphering the human genome. Although in recent months the leaders of the HGP announced that a “working draft” of the human Genome has been completed, the hope is to have a complete, error-free, final draft by 2003—coincidentally, the 50th anniversary of the discovery of DNA's molecular structure.

Summary of basic HGP goals:

- *Identify* all estimated 50,000-100,000 genes in human DNA
- *Determine sequence* of 3 billion chemical bases that make up human DNA
 - ***Human DNA sequence goals:***
 - Achieve *coverage* of at least 90% of Genome in *working draft* by the end of 2001—(moved up to spring 2000) - *Goal Reached* -
 - *Finish one-third* of the human Genome sequence by end of 2001
 - *Finish complete* human Genome sequence by end of 2003
 - Make sequence totally and freely accessible
- Create bioinformatics tools – Develop databases and analysis algorithms
- Store information in databases
- Develop faster, more efficient *sequencing technologies*
- Identify genes and coding regions – Develop efficient in-vitro or in-silico methods
- Develop tools for *data analysis*
- Map genomes of select *non-human* organisms
- Sequence other model organisms – Bacteria, yeast, fruit fly, worm, mouse
- Address *ethical, legal, and social issues* (ELSI) that may arise from project

Goals and Completion

Area	Goal	Achieved	Date Achieved
Genetic Map	2- to 5-cM resolution map (600 – 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < \$0.25 per finished base	Sequence >1,400 Mb/year at <\$0.09 per finished base	November 2002
Human Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene Identification	Full-length human cDNAs	15,000 full-length human cDNAs	March 2003
Model Organisms	Complete genome sequences of <i>E. coli</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i>	Finished genome sequences of <i>E. coli</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , plus whole-genome drafts of several others, including <i>C. briggsae</i> , <i>D. pseudoobscura</i> , mouse and rat	April 2003

Applications of Human Genome Project:

1. Better understanding of Polygenic disorders: The single gene disorders such as Cystic fibrosis, Sickle cell anemia are known. But many of the diseases such as Cancer, Hypertension are polygenic in nature. Sequencing of such genes helps us to better evaluate the disease giving more patient specific and friendly treatment.
2. Improvement in Gene Therapy: Genome sequencing helps in better provision of Gene therapy which is in its preliminary stage. This helps in effective treatment of genetic diseases.
3. Well elucidated Human genome sequence helps in improved diagnosis of many genetic disorders.
4. Development of Pharmacogenomics- Specialization in this field helps to know the individual genetic makeup thereby providing more personalized treatment.
5. Better cure of psychiatric disorders: Genes responsible for behavioral and psychiatric diseases can be better understood and treated.
6. An important application of HGP is better understanding of Mutations concept.

7. Better understanding of Developmental biology - Evolution from eggs to adults.
8. Human genome data also helps in development of Biotechnology in various spheres.

Findings

Large variation in GC content – Correlated with repeat content and gene density

- CpG dinucleotides are surprisingly rare – But CpG islands correlated with gene density
- Recombination rates are uneven – More recombination further from centromeres
- About 50% of genome is repeats – SINEs, LINEs, LTR retrotransposons, transposons

Mutation rates are uneven – Genome has more GC than equilibrium

- Differences between the sexes – Males mutate more but recombine less
- Many segmental duplications – 1–200 kb copied within or across chromosomes
- Estimated around 30,000 human genes – Unevenly distributed across chromosomes

MAJOR HIGHLIGHTS OF HGP :

1. Approximately 90pc of Human Genome was sequenced and the cause for underlying genetic disorders have been depicted.
2. The remaining 10pc is located at the end of chromosomes or at telomeres.
3. The human genome consisted of 3200 billion base pairs of which Gene and Gene Related sequences hold 1200 base pairs while Intergenic DNA contributed 2000 base pairs.
4. Proteins contribute 1.1-1.4 pc
5. Approximately 25% of the genome is composed of introns which appear as repeating units with no known functions.
6. Protein Coding Genes- 30000-40000
7. An average gene consists of 3000 bases. Dystrophin is the largest human gene with 2.4 million bases.
8. Chromosome 1 is the largest and contains 2968 genes while Chromosome Y is the smallest.
9. Genetic sequences that are associated with diseases like breast cancer, deafness, muscle diseases, blindness were sequenced and reported.
10. Repeated sequences constituted 50% of the genome.
11. 97% of the human genome has unknown functions
12. More than 3 million Single Nucleotide Polymorphisms have been identified.
13. Human DNA is 98% identical to Chimpanzees.
14. About 200 genes of human genome are found in bacteria too.

What is Next?

- Find all human genes – Only ~15,000 have yet been confirmed
- Identify effects of genetic variation – Understand diseases, healthy differences
- Understand non-coding regions – Chromosome structure, control mechanisms?
- Model human being as system – Within cells and as a whole organism

What were some of the ethical, legal, and social implications addressed by the Human Genome Project?

The Ethical, Legal, and Social Implications (ELSI) program was founded in 1990 as an integral part of the Human Genome Project. The mission of the ELSI program was to identify and address issues raised by genomic research that would affect individuals, families, and society. A percentage of the Human Genome Project budget at the National Institutes of Health and the U.S. Department of Energy was devoted to ELSI research.

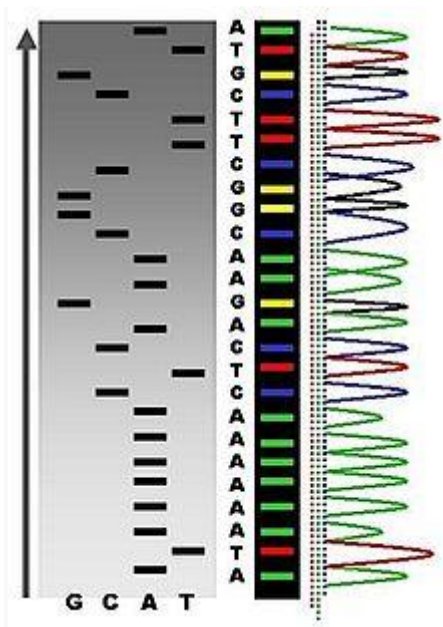
The ELSI program focused on the possible consequences of genomic research in four main areas:

- Privacy and fairness in the use of genetic information, including the potential for genetic discrimination in employment and insurance.
- The integration of new genetic technologies, such as genetic testing, into the practice of clinical medicine.
- Ethical issues surrounding the design and conduct of genetic research with people, including the process of informed consent.
- The education of healthcare professionals, policy makers, students, and the public about genetics and the complex issues that result from genomic research.

DNA sequencing

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as medical diagnosis, biotechnology, forensic biology, virology and biological systematics. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species.



An example of the results of automated chain-termination DNA sequencing.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of fluorescence-based sequencing methods with a DNA sequencer,^[1] DNA sequencing has become easier and orders of magnitude faster.

Uses of sequencing

DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e. clusters of genes or operons), full chromosomes or entire genomes. Sequencing provides the order of individual nucleotides present in molecules of DNA or RNA isolated from animals, plants, bacteria, archaea, or virtually any other source of genetic information.

This information is useful to various fields of biology and other sciences, medicine, forensics, and other areas of study.

Molecular biology

Sequencing is used in molecular biology to study genomes and the proteins they encode. Information obtained using sequencing allows researchers to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.

Evolutionary biology

Since DNA is an informative macromolecule in terms of transmission from one generation to another, DNA sequencing is used in evolutionary biology to study how different organisms are related and how they evolved.

Metagenomics

The field of metagenomics involves identification of organisms present in a body of water, sewage, dirt, debris filtered from the air, or swab samples from organisms. Knowing which organisms are present in a particular environment is critical to research in ecology, epidemiology, microbiology, and other fields. Sequencing enables researchers to determine which types of microbes may be present in a microbiome, for example.

Medicine

Medical technicians may sequence genes (or, theoretically, full genomes) from patients to determine if there is risk of genetic diseases. This is a form of genetic testing, though some genetic tests may not involve DNA sequencing.

Forensics

DNA sequencing may be used along with DNA profiling methods for forensic identification and paternity testing.

History

Deoxyribonucleic acid (DNA) was first discovered and isolated by Friedrich Miescher in 1869, but it remained understudied for many decades because proteins, rather than DNA, were thought to hold the genetic blueprint to life. This situation changed after 1944 as a result of some experiments by Oswald Avery, Colin MacLeod, and Maclyn McCarty demonstrated that purified DNA could change one strain of bacteria into another type. This was the first time that DNA was shown capable of transforming the properties of cells.

In 1953 James Watson and Francis Crick put forward their double-helix model of DNA which depicted DNA being made up of two strands of nucleotides coiled around each other, linked together by hydrogen bonds, in a spiral configuration. Each strand they argued was composed of four complementary nucleotides: adenine (A), cytosine (C), guanine (G) and

thymine (T) and was oriented in opposite directions. Such a structure they proposed allowed each strand to reconstruct the other and was central to the passing on of hereditary information between generations.^[7]

The foundation for sequencing DNA was first laid by the work of Fred Sanger who by 1955 had completed the sequence of all the amino acids in insulin, a small protein secreted by the pancreas. This provided the first conclusive evidence that proteins were chemical entities with a specific molecular pattern rather than a random mixture of material suspended in fluid. Sanger's success in sequencing insulin greatly electrified x-ray crystallographers, including Watson and Crick who by now were trying to understand how DNA directed the formation of proteins within a cell. Soon after attending a series of lectures given by Fred Sanger in October 1954, Crick began to develop a theory which argued that the arrangement of nucleotides in DNA determined the sequence of amino acids in proteins which in turn helped determine the function of a protein. He published this theory in 1958

RNA sequencing

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), in 1972^[9] and 1976.^[10]

Early DNA sequencing methods

The first method for determining DNA sequences involved a location-specific primer extension strategy established by Ray Wu at Cornell University in 1970.^[11] DNA polymerase catalysis and specific nucleotide labeling, both of which figure prominently in current sequencing schemes, were used to sequence the cohesive ends of lambda phage DNA.^{[12][13][14]} Between 1970 and 1973, Wu, R Padmanabhan and colleagues demonstrated that this method can be employed to determine any DNA sequence using synthetic location-specific primers.^{[15][16][17]} Frederick Sanger then adopted this primer-extension strategy to develop more rapid DNA sequencing methods at the MRC Centre, Cambridge, UK and published a method for "DNA sequencing with chain-terminating inhibitors" in 1977.^[18] Walter Gilbert and Allan Maxam at Harvard also developed sequencing methods, including one for "DNA sequencing by chemical degradation".^{[19][20]} In 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.^[21] Advancements in sequencing were aided by the concurrent development of recombinant DNA technology, allowing DNA samples to be isolated from sources other than viruses.

Sequencing of full genomes

The first full DNA genome to be sequenced was that of bacteriophage ϕ X174 in 1977.^[22] Medical Research Council scientists deciphered the complete DNA sequence of the Epstein-Barr virus in 1984, finding it contained 172,282 nucleotides. Completion of the sequence marked a significant turning point in DNA sequencing because it was achieved with no prior genetic profile knowledge of the virus.^[23]

A non-radioactive method for transferring the DNA molecules of sequencing reaction mixtures onto an immobilizing matrix during electrophoresis was developed by Pohl and co-workers in the early 1980s.^{[24][25]} Followed by the commercialization of the DNA sequencer "Direct-Blotting-Electrophoresis-System GATC 1500" by GATC Biotech, which was intensively used in the framework of the EU genome-sequencing programme, the complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome II.^[26] Leroy E. Hood's laboratory at the California Institute of Technology announced the first semi-automated DNA sequencing machine in 1986.^[27] This was followed by Applied Biosystems' marketing of the first fully automated sequencing machine, the ABI 370, in 1987 and by Dupont's Genesis 2000^[28] which used a novel fluorescent labeling technique enabling all four dideoxynucleotides to be identified in a single lane. By 1990, the U.S. National Institutes of Health (NIH) had begun large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* at a cost of US\$0.75 per base. Meanwhile, sequencing of human cDNA sequences called expressed sequence tags began in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.^[29] In 1995, Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) published the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*. The circular chromosome contains 1,830,137 bases and its publication in the journal Science^[30] marked the first published use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

By 2001, shotgun sequencing methods had been used to produce a draft sequence of the human genome

Next-generation sequencing methods

Several new methods for DNA sequencing were developed in the mid to late 1990s and were implemented in commercial DNA sequencers by the year 2000.

On October 26, 1990, Roger Tsien, Pepi Ross, Margaret Fahnestock and Allan J Johnston filed a patent describing stepwise ("base-by-base") sequencing with removable 3' blockers on DNA arrays (blots and single DNA molecules).^[33] In 1996, Pål Nyrén and his

student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm published their method of pyrosequencing.^[34]

On April 1, 1997, Pascal Mayer and Laurent Farinelli submitted patents to the World Intellectual Property Organization describing DNA colony sequencing.^[35] The DNA sample preparation and random surface-PCR arraying methods described in this patent, coupled to Roger Tsien et al.'s "base-by-base" sequencing method, is now implemented in Illumina's Hi-Seq genome sequencers.

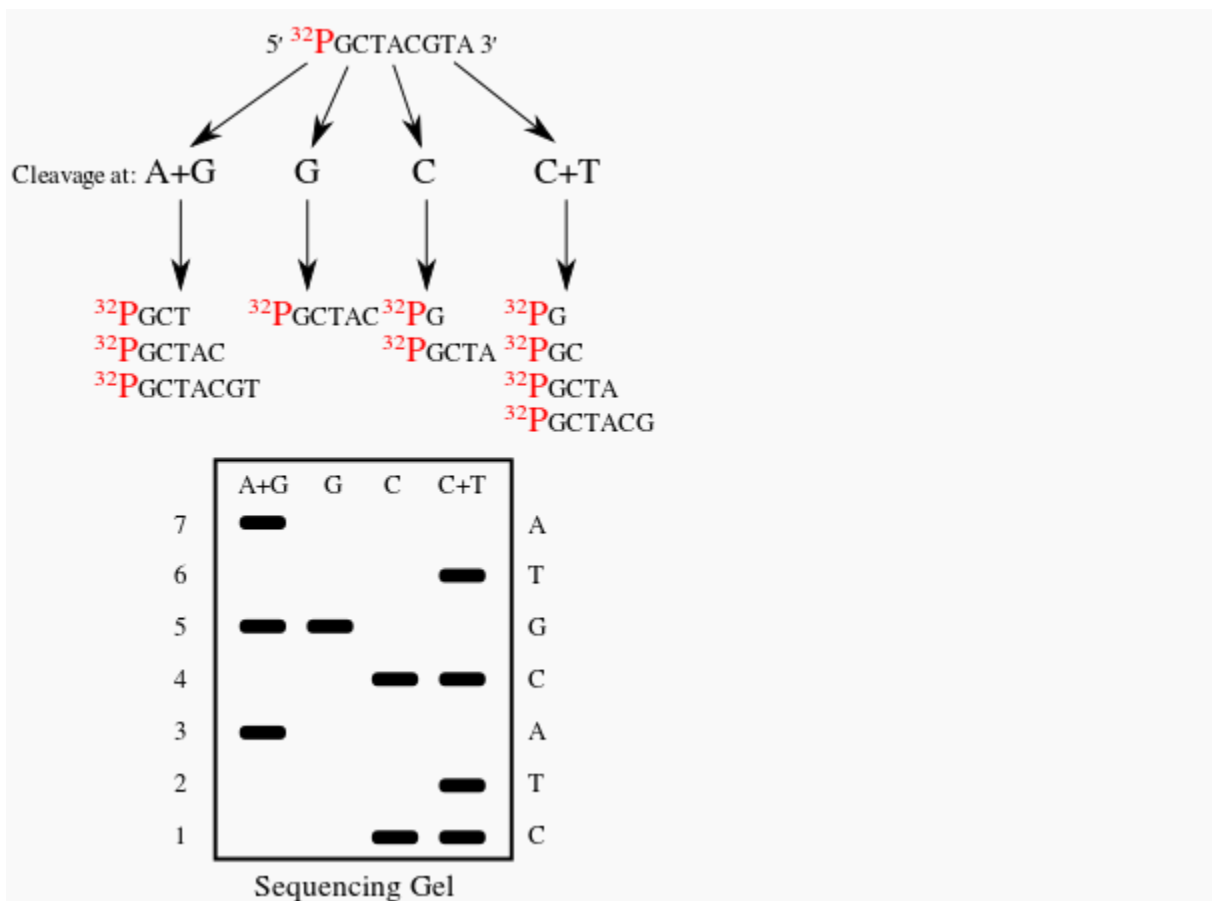
Lynx Therapeutics published and marketed "Massively parallel signature sequencing", or MPSS, in 2000. This method incorporated a parallelized, adapter/ligation-mediated, bead-based sequencing technology and served as the first commercially available "next-generation" sequencing method, though no DNA sequencers were sold to independent laboratories.^[36]

In 2004, 454 Life Sciences marketed a parallelized version of pyrosequencing.^[37] The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of the new generation of sequencing technologies, after MPSS.^[38]

The large quantities of data produced by DNA sequencing have also required development of new methods and programs for sequence analysis. Phil Green and Brent Ewing of the University of Washington described their phred quality score for sequencer data analysis in 1998

Maxam–Gilbert sequencing

Maxam–Gilbert sequencing is a method of DNA sequencing developed by Allan Maxam and Walter Gilbert in 1976–1977. This method is based on nucleobase-specific partial chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides.^[1]



An example Maxam–Gilbert sequencing reaction. Cleaving the same tagged segment of DNA at different points yields tagged fragments of different sizes. The fragments may then be separated by gel electrophoresis.

Maxam–Gilbert sequencing was the first widely adopted method for DNA sequencing, and, along with the Sanger dideoxy method, represents the first generation of DNA sequencing methods. Maxam–Gilbert sequencing is no longer in widespread use, having been supplanted by next-generation sequencing.

Procedure

Maxam–Gilbert sequencing requires radioactive labeling at one 5' end of the DNA fragment to be sequenced (typically by a kinase reaction using gamma-³²P ATP) and purification of the DNA. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are hydrolysed using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the reaction of thymine for the C-only reaction. The modified DNAs may then be cleaved by hot piperidine; (CH₂)₅NH at the position of the modified base. The concentration of the

modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule.

The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each showing the location of identical radiolabeled DNA molecules. From presence and absence of certain fragments the sequence may be inferred

Sanger sequencing

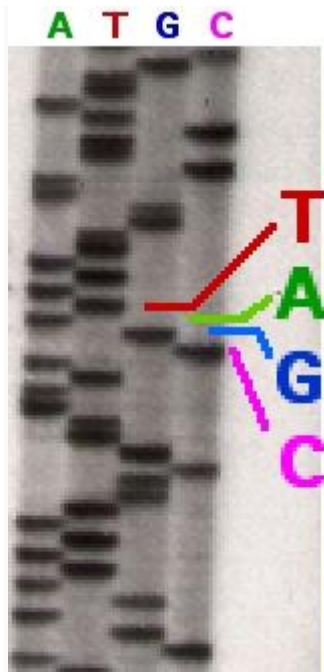
Sanger sequencing is a method of **DNA sequencing** based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.^{[1][2]} Developed by Frederick Sanger and colleagues in 1977, it was the most widely used sequencing method for approximately 25 years. More recently, Sanger sequencing has been supplanted by "Next-Gen" sequencing methods, especially for large-scale, automated genome analyses. However, the Sanger method remains in wide use, for smaller-scale projects, validation of Next-Gen results and for obtaining especially long contiguous DNA sequence reads (>500 nucleotides).

Method

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleosidetriphosphates (dNTPs), and modified di-deoxynucleosidetriphosphates (ddNTPs), the latter of which terminate DNA strand elongation. These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. The ddNTPs may be radioactively or fluorescently labeled for detection in automated sequencing machines.

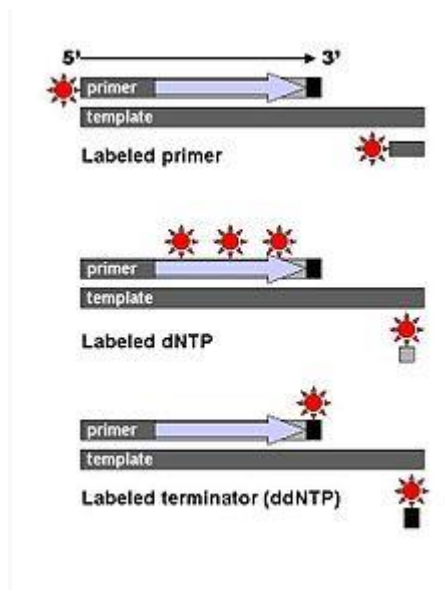
The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP), while the other added nucleotides are ordinary ones. The dideoxynucleotide is added in approximately 100-fold excess of the corresponding deoxynucleotide (e.g. 0.005mM dATP : 0.5mM ddATP) allowing for enough fragments to be produced while still transcribing the complete sequence.^[2] Putting it in a more sensible order, four separate reactions are needed in this process to test all four ddNTPs. Following rounds of template DNA extension

from the bound primer, the resulting DNA fragments are heat denatured and separated by size using gel electrophoresis. In the original publication of 1977,^[2] the formation of base-paired loops of ssDNA was a cause of serious difficulty in resolving bands at some locations. This is frequently performed using a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C). The DNA bands may then be visualized by autoradiography or UV light and the DNA sequence can be directly read off the X-ray film or gel image.



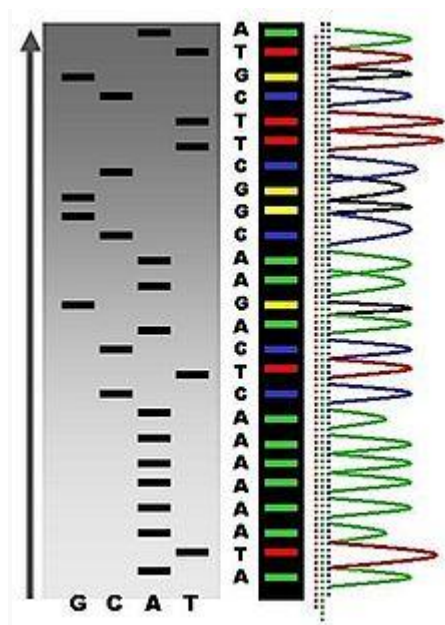
Part of a radioactively labelled sequencing gel

In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes, from bottom to top, are then used to read the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers^{[3][4]} of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.

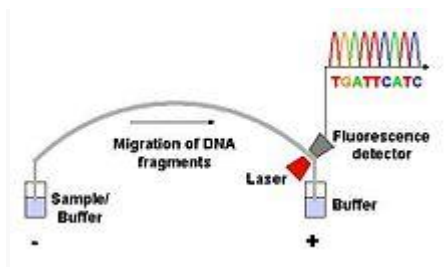


Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for

sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



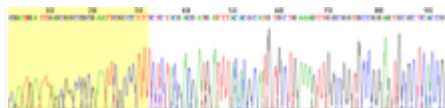
Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis (see figure to the left).

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch. Batch runs may occur up to 24 times a day. DNA sequencers separate strands by size (or length) using capillary electrophoresis, they detect and record dye fluorescence, and output data as fluorescent peak trace chromatograms. Sequencing reactions (thermocycling and labelling), cleanup and re-suspension of samples in a buffer solution are performed separately, before loading samples onto the sequencer. A number of commercial

and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (which are generally located at the ends of the sequence). The accuracy of such algorithms is inferior to visual examination by a human operator, but is adequate for automated processing of large sequence data sets.

Challenges

Common challenges of DNA sequencing with the Sanger method include poor quality in the first 15-40 bases of the sequence due to primer binding and deteriorating quality of sequencing traces after 700-900 bases. Base calling software such as Phred typically provides an estimate of quality to aid in trimming of low-quality regions of sequences.^{[5][6]}

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and next-generation sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.^{[7][8]}

Current methods can directly sequence only relatively short (300-1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide.

Microfluidic Sanger sequencing

Microfluidic Sanger sequencing is a lab-on-a-chip application for DNA sequencing, in which the Sanger sequencing steps (thermal cycling, sample purification, and capillary electrophoresis) are integrated on a wafer-scale chip using nanoliter-scale sample volumes. This technology generates long and accurate sequence reads, while obviating many of the significant shortcomings of the conventional Sanger method (e.g. high consumption of expensive reagents, reliance on expensive equipment, personnel-intensive manipulations, etc.) by integrating and automating the Sanger sequencing steps.

In its modern inception, high-throughput genome sequencing involves fragmenting the genome into small single-stranded pieces, followed by amplification of the fragments by Polymerase Chain Reaction (PCR). Adopting the Sanger method, each DNA fragment is irreversibly terminated with the incorporation of a fluorescently labeled dideoxy chain-terminating nucleotide, thereby producing a DNA “ladder” of fragments that each differ in length by one base and bear a base-specific fluorescent label at the terminal base. Amplified

base ladders are then separated by Capillary Array Electrophoresis (CAE) with automated, *in situ* “finish-line” detection of the fluorescently labeled ssDNA fragments, which provides an ordered sequence of the fragments. These sequence reads are then computer assembled into overlapping or contiguous sequences (termed “contigs”) which resemble the full genomic sequence once fully assembled.^[9]

Sanger methods achieve read lengths of approximately 800bp (typically 500-600bp with non-enriched DNA). The longer read lengths in Sanger methods display significant advantages over other sequencing methods especially in terms of sequencing repetitive regions of the genome. A challenge of short-read sequence data is particularly an issue in sequencing new genomes (*de novo*) and in sequencing highly rearranged genome segments, typically those seen of cancer genomes or in regions of chromosomes that exhibit structural variation.^[10]

Applications of microfluidic sequencing technologies

Other useful applications of DNA sequencing include single nucleotide polymorphism (SNP) detection, single-strand conformation polymorphism (SSCP) heteroduplex analysis, and short tandem repeat (STR) analysis. Resolving DNA fragments according to differences in size and/or conformation is the most critical step in studying these features of the genome.^[9]

Device design

The sequencing chip has a four-layer construction, consisting of three 100-mm-diameter glass wafers (on which device elements are microfabricated) and a polydimethylsiloxane (PDMS) membrane. Reaction chambers and capillary electrophoresis channels are etched between the top two glass wafers, which are thermally bonded. Three-dimensional channel interconnections and microvalves are formed by the PDMS and bottom manifold glass wafer. The device consists of three functional units, each corresponding to the Sanger sequencing steps. The Thermal Cycling (TC) unit is a 250-nanoliter reaction chamber with integrated resistive temperature detector, microvalves, and a surface heater. Movement of reagent between the top all-glass layer and the lower glass-PDMS layer occurs through 500- μ m-diameter via-holes. After thermal-cycling, the reaction mixture undergoes purification in the capture/purification chamber, and then is injected into the capillary electrophoresis (CE) chamber. The CE unit consists of a 30-cm capillary which is folded into a compact switchback pattern via 65- μ m-wide turns.

Sequencing chemistry

- **Thermal cycling**

In the TC reaction chamber, dye-terminator sequencing reagent, template DNA, and primers are loaded into the TC chamber and thermal-cycled for 35 cycles (at 95°C for 12 seconds and at 60°C for 55 seconds).

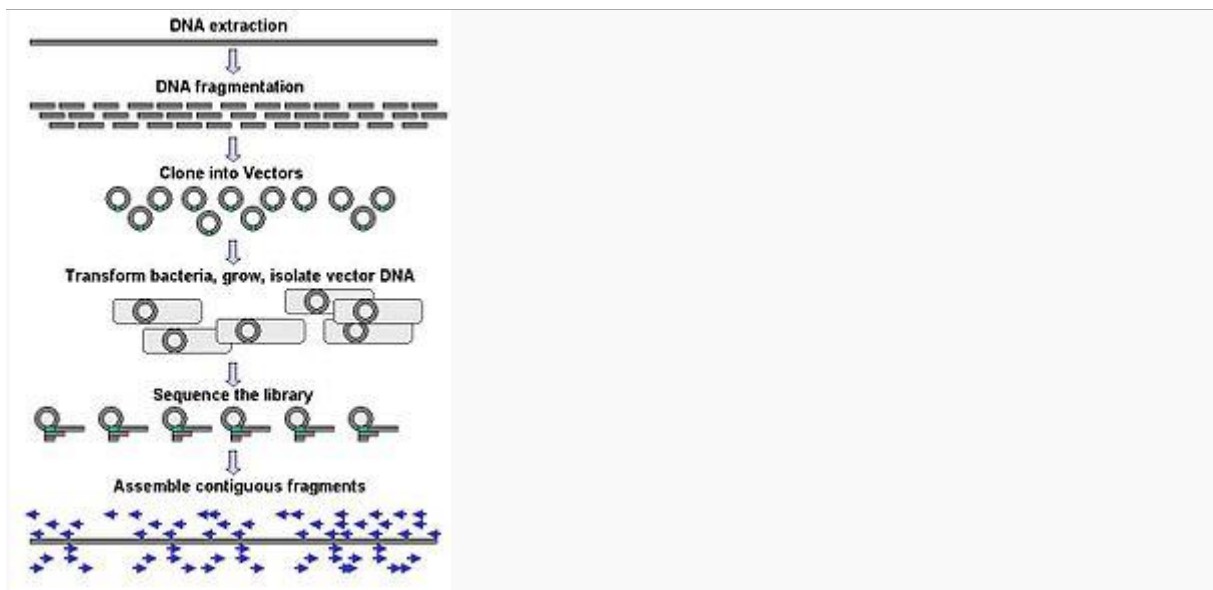
- **Purification**

The charged reaction mixture (containing extension fragments, template DNA, and excess sequencing reagent) is conducted through a capture/purification chamber at 30°C via a 33-Volts/cm electric field applied between capture outlet and inlet ports. The capture gel through which the sample is driven, consists of 40 μ M of oligonucleotide (complementary to the primers) covalently bound to a polyacrylamide matrix. Extension fragments are immobilized by the gel matrix, and excess primer, template, free nucleotides, and salts are eluted through the capture waste port. The capture gel is heated to 67-75°C to release extension fragments.

- **Capillary electrophoresis**

Extension fragments are injected into the CE chamber where they are electrophoresed through a 125-167-V/cm field.

Advanced methods and *de novo* sequencing



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions.(click to expand)

Large-scale sequencing often aims at sequencing very long DNA pieces, such as whole chromosomes, although large-scale sequencing can also be used to generate very large numbers of short sequences, such as found in phage display. For longer targets such as chromosomes, common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented

DNA may then be cloned into a DNA vector and amplified in a bacterial host such as *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence. Studies have shown that adding a size selection step to collect DNA fragments of uniform size can improve sequencing efficiency and accuracy of the genome assembly. In these studies, automated sizing has proven to be more reproducible and precise than manual gel sizing.^{[42][43][44]}

The term "*de novo* sequencing" specifically refers to methods used to determine the sequence of DNA with no previously known sequence. *De novo* translates from Latin as "from the beginning". Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; shotgun methods are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR^[45] isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. A polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods developed by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, later Applied Biosystems, now Life Technologies).^{[46][47][48]}

Shotgun sequencing

In genetics, **shotgun sequencing**, also known as **shotgun cloning**, is a method used for sequencing long DNA strands. It is named by analogy with the rapidly expanding, quasi-random firing pattern of a shotgun.

The chain termination method of DNA sequencing (or "Sanger sequencing" for its developer Frederick Sanger) can only be used for fairly short strands of 100 to 1000 base pairs. Longer sequences are subdivided into smaller fragments that can be sequenced separately, and subsequently they are re-assembled to give the overall sequence. Two principal methods are used for this: primer walking (or "chromosome walking") which progresses through the entire strand piece by piece, and shotgun sequencing which is a faster but more complex process that uses random fragments.

In shotgun sequencing,^{[1][2]} DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain *reads*. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence.^[1]

Shotgun sequencing was one of the precursor technologies that was responsible for enabling full genome sequencing

Whole genome shotgun sequencing

Whole genome shotgun sequencing for small (4000- to 7000-base-pair) genomes was already in use in 1979.^[1] Broader application benefited from pairwise end sequencing, known colloquially as *double-barrel shotgun sequencing*. As sequencing projects began to take on longer and more complicated DNA sequences, multiple groups began to realize that useful information could be obtained by sequencing both ends of a fragment of DNA. Although sequencing both ends of the same fragment and keeping track of the paired data was more cumbersome than sequencing a single end of two distinct fragments, the knowledge that the two sequences were oriented in opposite directions and were about the length of a fragment apart from each other was valuable in reconstructing the sequence of the original target fragment. The first published description of the use of paired ends was in 1990^[4] as part of the sequencing of the human HGPRT locus, although the use of paired ends was limited to closing gaps after the application of a traditional shotgun sequencing approach. The first theoretical description of a pure pairwise end sequencing strategy, assuming fragments of constant length, was in 1991.^[5] At the time, there was community consensus that the optimal fragment length for pairwise end sequencing would be three times the sequence read length. In 1995 Roach et al.^[6] introduced the innovation of using fragments of varying sizes, and demonstrated that a pure pairwise end-sequencing strategy would be possible on large targets. The strategy was subsequently adopted by The Institute for Genomic Research (TIGR) to sequence the genome of the bacterium *Haemophilus influenzae* in 1995,^[7] and then by Celera Genomics to sequence the *Drosophila melanogaster* (fruit fly) genome in 2000,^[8] and subsequently the human genome.

To apply the strategy, a high-molecular-weight DNA strand is sheared into random fragments, size-selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends using the chain termination method yielding two short sequences. Each sequence is called an *end-read* or *read* and two reads from the same clone are referred to as *mate pairs*. Since the chain termination method

usually can only produce reads between 500 and 1000 bases long, in all but the smallest clones, mate pairs will rarely overlap.

The original sequence is reconstructed from the reads using sequence assembly software. First, overlapping reads are collected into longer composite sequences known as *contigs*. Contigs can be linked together into *scaffolds* by following connections between mate pairs. The distance between contigs can be inferred from the mate pair positions if the average fragment length of the library is known and has a narrow window of deviation. Depending on the size of the gap between contigs, different techniques can be used to find the sequence in the gaps. If the gap is small (5-20kb) then the use of PCR to amplify the region is required, followed by sequencing. If the gap is large (>20kb) then the large fragment is cloned in special vectors such as BAC (Bacterial artificial chromosomes) followed by sequencing of the vector.

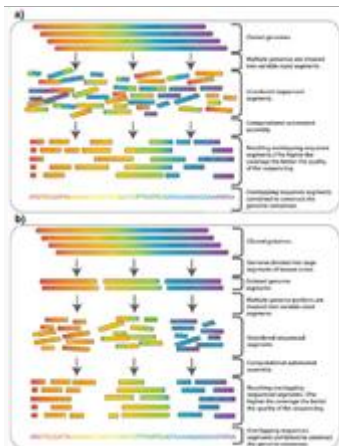
Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than more traditional approaches. Detractors argue that although the technique quickly sequences large regions of DNA, its ability to correctly link these regions is suspect, particularly for genomes with repeating regions. As sequence assembly programs become more sophisticated and computing power becomes cheaper, it may be possible to overcome this limitation.^[citation needed]

Coverage

Coverage (read depth or depth) is the average number of reads representing a given nucleotide in the reconstructed sequence. It can be calculated from the length of the original genome (G), the number of reads (N), and the average read length (L) as $N \times L / G$. For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2x redundancy. This parameter also enables one to estimate other quantities, such as the percentage of the genome covered by reads (sometimes also called coverage). A high coverage in shotgun sequencing is desired because it can overcome errors in base calling and assembly. The subject of DNA sequencing theory addresses the relationships of such quantities.

Sometimes a distinction is made between *sequence coverage* and *physical coverage*. Sequence coverage is the average number of times a base is read (as described above). Physical coverage is the average number of times a base is read or spanned by mate paired reads.

Hierarchical Shotgun sequencing

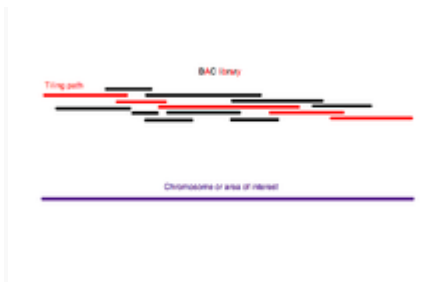


In whole genome shotgun sequencing (top), the entire genome is sheared randomly into small fragments (appropriately sized for sequencing) and then reassembled. In hierarchical shotgun sequencing (bottom), the genome is first broken into larger segments. After the order of these segments is deduced, they are further sheared into fragments appropriately sized for sequencing.

Although shotgun sequencing can in theory be applied to a genome of any size, its direct application to the sequencing of large genomes (for instance, the Human Genome) was limited until the late 1990s, when technological advances made practical the handling of the vast quantities of complex data involved in the process.^[10] Historically, full-genome shotgun sequencing was believed to be limited by both the sheer size of large genomes and by the complexity added by the high percentage of repetitive DNA (greater than 50% for the human genome) present in large genomes.^[11] It was not widely accepted that a full-genome shotgun sequence of a large genome would provide reliable data. For these reasons, other strategies that lowered the computational load of sequence assembly had to be utilized before shotgun sequencing was performed.^[11] In hierarchical sequencing, also known as top-down sequencing, a low-resolution physical map of the genome is made prior to actual sequencing. From this map, a minimal number of fragments that cover the entire chromosome are selected for sequencing.^[12] In this way, the minimum amount of high-throughput sequencing and assembly is required.

The amplified genome is first sheared into larger pieces (50-200kb) and cloned into a bacterial host using BACs or PACs. Because multiple genome copies have been sheared at random, the fragments contained in these clones have different ends, and with enough

coverage (see section above) finding a **scaffold** of BAC contigs that covers the entire genome is theoretically possible. This scaffold is called a **tiling path**.



A BAC contig that covers the entire genomic area of interest makes up the tiling path.

Once a tiling path has been found, the BACs that form this path are sheared at random into smaller fragments and can be sequenced using the shotgun method on a smaller scale.

Although the full sequences of the BAC contigs is not known, their orientations relative to one another are known. There are several methods for deducing this order and selecting the BACs that make up a tiling path. The general strategy involves identifying the positions of the clones relative to one another and then selecting the least number of clones required to form a contiguous scaffold that covers the entire area of interest. The order of the clones is deduced by determining the way in which they overlap.^[13] Overlapping clones can be identified in several ways. A small radioactively or chemically labeled probe containing a sequence-tagged site (STS) can be hybridized onto a microarray upon which the clones are printed.^[13] In this way, all the clones that contain a particular sequence in the genome are identified. The end of one of these clones can then be sequenced to yield a new probe and the process repeated in a method called chromosome walking.

Alternatively, the BAC library can be restriction-digested. Two clones that have several fragment sizes in common are inferred to overlap because they contain multiple similarly spaced restriction sites in common.^[13] This method of genomic mapping is called restriction fingerprinting because it identifies a set of restriction sites contained in each clone. Once the overlap between the clones has been found and their order relative to the genome known, a scaffold of a minimal subset of these contigs that covers the entire genome is shotgun-sequenced.^[12]

Because it involves first creating a low-resolution map of the genome, hierarchical shotgun sequencing is slower than whole-genome shotgun sequencing, but relies less heavily on computer algorithms than whole-genome shotgun sequencing. The process of extensive BAC library creation and tiling path selection, however, make hierarchical shotgun sequencing slow and labor-intensive. Now that the technology is available and the reliability of the data

demonstrated,^[11] and the speed and cost efficiency of whole-genome shotgun sequencing has made it the primary method for genome sequencing.

Shotgun and Next-generation sequencing

The classical shotgun sequencing was based on the Sanger sequencing method: this was the most advanced technique for sequencing genomes from about 1995–2005. The shotgun strategy is still applied today, however using other sequencing technologies, called next-generation sequencing. These technologies produce shorter reads (anywhere from 25–500bp) but many hundreds of thousands or millions of reads in a relatively short time (on the order of a day).^[14] This results in high coverage, but the assembly process is much more computationally intensive. These technologies are vastly superior to Sanger sequencing due to the high volume of data and the relatively short time it takes to sequence a whole genome

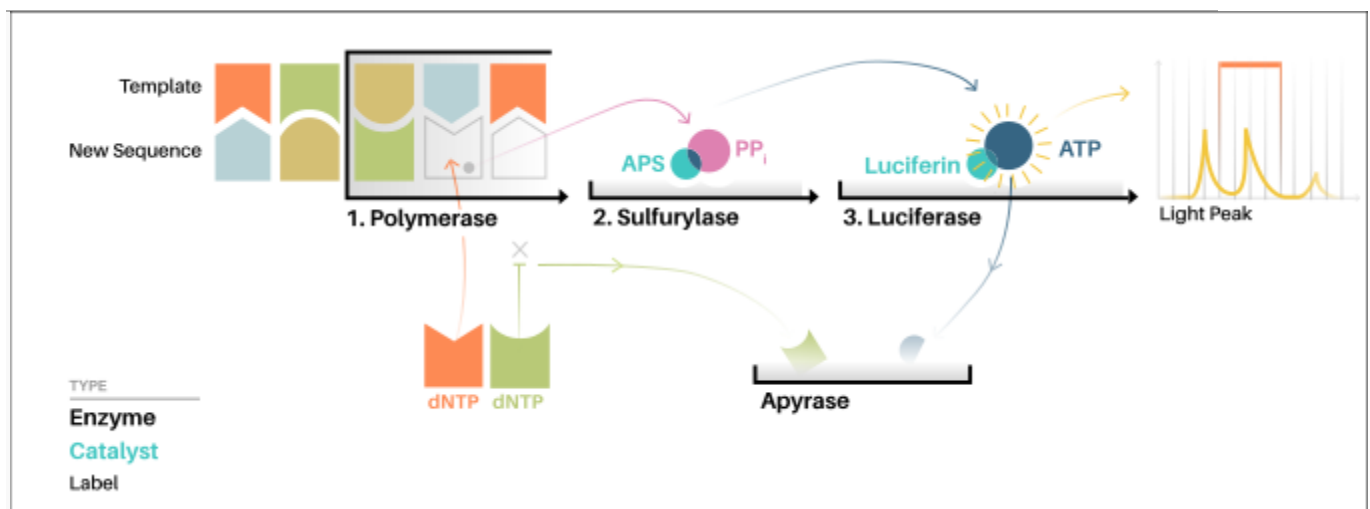
Bridge PCR

Another method for *in vitro* clonal amplification is bridge PCR, in which fragments are amplified upon primers attached to a solid surface^{[35][50][51]} and form "DNA colonies" or "DNA clusters". This method is used in the Illumina Genome Analyzer sequencers. Single-molecule methods, such as that developed by Stephen Quake's laboratory (later commercialized by Helicos) are an exception: they use bright fluorophores and laser excitation to detect base addition events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.^[52]

Pyrosequencing

Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA) based on the "sequencing by synthesis" principle. It differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides.^[1] The technique was developed by Mostafa Ronaghi and Pål Nyrén at the Royal Institute of Technology in Stockholm in 1996.^{[2][3][4]} The desired DNA sequence is able to be determined by light emitted upon incorporation of the next complementary nucleotide by the fact that only one out of four of the possible A/T/C/G nucleotides are added and available at a time so that only one letter can be incorporated on the single stranded template (which is the sequence to be determined). The intensity of the light determines if there are more than one of these "letters" in a row. The previous nucleotide letter (one out of four possible dNTP) is degraded before the next nucleotide letter is added for synthesis: allowing for the possible revealing of the next nucleotide(s) via the resulting intensity of light (if the nucleotide added was the next complementary letter in the sequence). This process is repeated with each of the four letters until the DNA sequence of the single stranded template is determined.

Procedure



The chart shows how pyrosequencing works.

"Sequencing by synthesis" involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The pyrosequencing method is based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time,

and detecting which base was actually added at each step. The template DNA is immobile, and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template.

The single-strand DNA (ssDNA) template is hybridized to a sequencing primer and incubated with the enzymes DNA polymerase, ATP sulfurylase, luciferase and apyrase, and with the substrates adenosine 5' phosphosulfate (APS) and luciferin.

1. The addition of one of the four deoxynucleoside triphosphates (dNTPs) (dATP α S, which is not a substrate for a luciferase, is added instead of dATP to avoid noise) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi).
2. ATP sulfurylase converts PPi to ATP in the presence of adenosine 5' phosphosulfate. This ATP acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a pyrogram.
3. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide.

Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA. As of 2007, pyrosequencing is most commonly used for resequencing or sequencing of genomes for which the sequence of a close relative is already available.

The templates for pyrosequencing can be made both by solid phase template preparation (streptavidin-coated magnetic beads) and enzymatic template preparation (apyrase+exonuclease). So Pyrosequencing can be differentiated into two types, namely Solid Phase Pyrosequencing and Liquid Phase Pyrosequencing.

Commercialization

The company Pyrosequencing AB in Uppsala, Sweden was founded with venture capital provided by HealthCap in order to commercialize machinery and reagents for sequencing short stretches of DNA using the pyrosequencing technique. Pyrosequencing AB was listed on the Stockholm Stock Exchange in 1999. It was renamed to Biotage in 2003. The

pyrosequencing business line was acquired by Qiagen in 2008.^[5] Pyrosequencing technology was further licensed to 454 Life Sciences. 454 developed an array-based pyrosequencing technology which has emerged as a platform for large-scale DNA sequencing. Most notable are the applications for genome sequencing and metagenomics. *GS FLX*, the latest pyrosequencing platform by 454 Life Sciences (now owned by Roche Diagnostics), can generate 400 Mb in a 10-hour run with a single machine. Each run would cost about 5,000-7,000 USD.

Genome databases

A genome comprises all of the genetic material in the chromosomes of a particular organism. Genome databases are an organized collection of information that have resulted from the production or mapping of genome (sequence) or genome product (transcript, protein) information. These databases collect genome sequences, annotate and analyze them, and provide public access. Some add curation of experimental literature to improve computed annotations. These databases may hold many species genomes, or a single model organism genome.

Human Genome Databases, Browsers and Variation Resources

- Database of Genomic Variants
- dbVar Database of Genomic Structural Variation
- ENCODE Project ENCyclopedia Of DNA Elements
- Ensembl Human human genes generated automatically by the Ensembl gene builder
- Entrez Gene searchable database of genes, defined by sequence and/or located in the NCBI Map Viewer
- Genome Reference Consortium Putting sequences into a chromosome context
- GWAS Central centralized compilation of summary level findings from genetic association studies
- HapMap international HapMap Project
- H-Invitational Database an integrated database of human genes and transcripts
- Human Genome Segmental Duplication Database
- Human Structural Variation Database
- 1000 Genomes A Deep Catalog of Human Genetic Variation
- UCSC Human Genome Browser Gateway
- VEGA Human manual annotation of finished genome sequence

Other Vertebrate Genome Databases and Browsers

- AgBase a curated, open-source resource for functional analysis of agricultural plant and animal gene products
- AnolisGenome a community resource site for Anolis genomics and genetic studies
- ARKdb species databases includes: Cat, Chicken, Cow, Deer, Horse, Pig, Salmon, Sheep, Tilapia, Turkey
- BirdBase A Database of Avian Genes and Genomes
- Bovmap mapping the Bovine genome
- Lyons Feline & Comparative Genetics
- Chicken Genome Resources
- The Dog Genome Project
- Ensembl genome databases for vertebrates and other eukaryotic species
- Entrez Gene searchable database of genes, from RefSeq genomes, defined by sequence and/or located in the NCBI Map Viewer
- Fugu the Fugu genomics project
- Horse Genome Project
- Kangaroo Genome Project
- lizardbase a centralized and consolidated informatics resource for lizard research
- MGI Mouse Genome Informatics
- National Animal Genome Research Program
- Pig Genome Coordination Program
- Porcine Genome Sequencing Project
- Pig Genome Resources
- Rabbit Genome Resources
- RGD Rat Genome Database
- Tetraodon Genome Browser
- UCSC Genome Bioinformatics
- VEGA Vertebrate Genome Annotation containing manual annotation of vertebrate finished genome sequence
- Xenbase a Xenopus web resource
- ZFIN Zebrafish Information Network

Non-Vertebrate Genome Databases and Browsers

- ANISEED Ascidian Network for InSitu Expression and Embryological Data
- AspGDA *Aspergillus* Genome Database
- BeetleBase the model organism database for *Tribolium castaneum*

- Cacao Genome Database
- *Caenorhabditis* Genome Sequencing Projects
- *Candida* Genome Database
- Chlamyddb database for the green alga *Chlamydomonas reinhardtii* and related species
- The Cotton Genome Database
- Daphnia Genome Database
- Dendrome A Forest Tree Genome Database
- dictyBase central resource for Dictyostelid genomics
- EcoGene the Database of *Escherichia coli* Sequence and Function
- Ensembl Genomes
- FlyBase a database of the *Drosophila* genome
- GenProtEC *E.Coli* genome and proteome database
- GOBASE the Organelle Genome Database
- Gramene a resource for comparative grass genomics
- HGD Hymenoptera Genome Database
- IGGI International Glossina Genome Initiative
- PomBase a scientific resource for fission yeast
- SGD *Saccharomyces* Genome Database
- SpBase *Strongylocentrotus purpuratus* Sea Urchin Genome Database
- StellaBase *Nematostella vectensis* Genomics Database
- TAIR The Arabidopsis Information Resource
- VectorBase invertebrate vectors of human pathogens
- WormBase the biology and genome of *C. elegans*

Proteomics Databases

- Proteomics Identifications Database (PRIDE) A public repository for proteomics data, containing protein and peptide identifications and their associated supporting evidence as well as details of post-translational modifications. (European Bioinformatics Institute)
- ProteomeXchange provides a coordinated submission of mass spectrometry proteomics data to the main existing proteomics repositories. It includes datasets such as PRIDE, Tranche, and PeptideAtlas.
- jPOSTrepo Japanese ProteOme STandard repository

- ProteomeScout - A public repository of processed proteomics datasets concerning post-translational modifications, includes quantification across conditions (if applicable). Also includes a graphics exports of protein annotations.
- MitoMiner - A mitochondrial proteomics database integrating large-scale experimental datasets from mass spectrometry and GFP studies for 12 species. (MRC Mitochondrial Biology Unit)
- GelMap - A public database of proteins identified on 2D gels (University of Hanover Proteomics Department)
- OWL - A public non-redundant database for protein search, derived from : SWISS PROT, PIR, GenBank(translation) and NRL-3D
- Proteome-pI pre-computed isoelectric points for >5000 proteomes of model organisms

PPI databases

The primary resources for PPI data are individual scientific publications. Several public databases collect published PPI data and provide researchers access to their curated datasets. These usually reference the original publication and the experimental method that determined every individual interaction. Database designers choose to represent these data in different ways, and the wide spectrum of experimental methods makes it difficult to design a single data model to capture all necessary experimental detail. To overcome this problem, the International Molecular Exchange (IMEx; <http://imex.sourceforge.net/>) consortium was formed. IMEx aims to enable the exchange of data and to avoid the duplication of the curation effort. To that end, an XML-based proteomics standard, termed the proteomics standards initiative - molecular interaction (PSI-MI) has been developed [17]. At the time of writing, however, no data had yet been exchanged, and it was therefore necessary to combine PPI data from all available databases using the authors' own scripts to obtain as comprehensive a network as possible.

Here, the focus is on six databases: the Biological General Repository for Interaction Datasets (BioGRID) [18], the Molecular INTeraction database (MINT) [19], the Biomolecular Interaction Network Database (BIND) [20], the Database of Interacting Proteins (DIP) [21], the IntAct molecular interaction database (IntAct)[22] and the Human Protein Reference Database (HPRD)[23] (see Table 1). These databases report only experimentally verified interactions.

PPI databases

Database	URL	Proteins	Interactions	Publications	Organisms
BioGRID	http://www.thebiogrid.org	23,341	90,972	16,369	10
MINT	http://mint.bio.uniroma2.it/mint	27,306	80,039	3,047	144
BIND	http://bond.unleashedinformatics.com	23,643	43,050	6,364	80
DIP	http://dip.doe-mbi.ucla.edu	21,167	53,431	3,193	134
IntAct	http://www.ebi.ac.uk/intact	37,904	129,559	3,166	131
HPRD	http://www.hprd.org	9,182	36,169	18,777	1

DIP, IntAct and MINT are active members of the IMEx initiative; the curation accuracy of these three databases was assessed recently by Cusick *et al.* [24] HPRD focuses entirely on human proteins, providing not only information on protein interactions, but also a variety of protein-specific information, such as post-translational modifications, disease associations and enzyme-substrate relationships. One of the first interaction databases, BIND, initiated in 2001 by the University of Toronto and the University of British Columbia, is part of the Biomolecular Object Network Databank (BOND) and was subsequently acquired by the company Thomson Reuters.

The following comparison is based on complete sets of binary interactions that were downloaded from the individual databases in May 2008. IntAct and MINT derive binary interactions from protein complexes using the spokes model. No other database provided any information on which model is applied. Only 'physical interactions' are considered here, although most databases also provide 'genetic interactions' -- that is, two non-essential genes that lead to a non-viable phenotype if they are knocked out simultaneously. Furthermore, interactions were only accepted if a publication identifier was provided along with the interacting proteins.

Currently, the most comprehensive database in terms of individual interactions is IntAct, with almost 130,000 unique interactions from up to 131 different organisms. Despite these large numbers, it cites only about 3,000 different publications. Whereas IntAct seems to be concentrating on high-throughput studies, HPRD also takes into account small-scale publications. Although being restricted to human proteins, it reports over 36,000 unique interactions from more than 18,000 publications. Only BioGRID cites a similar number of publications (16,369); it is also the second largest database in terms of the number of unique interactions. It should be noted that the databases examine publications in different depth, and that higher numbers of publications do not necessarily involve a higher curation effort.

The majority of known protein interactions account for proteins from *Saccharomyces cerevisiae* and *Homo sapiens*. Individual high-throughput interaction screens were carried out for some other organisms; these high-throughput studies usually account for the majority of all known interactions in the corresponding organism. By contrast, known protein interactions for *S. cerevisiae* and *H. sapiens* are dispersed over numerous publications. For this reason, the number of interactions for humans and yeast can vary considerably between different databases, depending on their coverage of the literature.



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY**

UNIT – IV - Fundamentals of Genomics and Proteomics– SBI1309

Protein purification

Protein purification is a series of processes intended to isolate one or a few proteins from a complex mixture, usually cells, tissues or whole organisms. Protein purification is vital for the characterization of the function, structure and interactions of the protein of interest. The purification process may separate the protein and non-protein parts of the mixture, and finally separate the desired protein from all other proteins. Separation of one protein from all others is typically the most laborious aspect of protein purification. Separation steps usually exploit differences in protein size, physico-chemical properties, binding affinity and biological activity. The pure result may be termed **protein isolate**.

Purpose

Protein purification is either preparative or analytical. **Preparative purifications** aim to produce a relatively large quantity of purified proteins for subsequent use. Examples include the preparation of commercial products such as enzymes (e.g. lactase), nutritional proteins (e.g. soy protein isolate), and certain biopharmaceuticals (e.g. insulin). **Analytical purification** produces a relatively small amount of a protein for a variety of research or analytical purposes, including identification, quantification, and studies of the protein's structure, post-translational modifications and function. Pepsin and urease were the first proteins purified to the point that they could be crystallized.^[1]



Recombinant bacteria can be grown in a flask containing growth media.

Preliminary steps

Extraction

If the protein of interest is not secreted by the organism into the surrounding solution, the first step of each purification process is the disruption of the cells containing the protein. Depending on how fragile the protein is and how stable the cells are, one could, for instance, use one of the following methods: i) repeated freezing and thawing, ii) sonication, iii) homogenization by high pressure (French press), iv) homogenization by grinding (bead mill), and v) permeabilization by detergents (e.g. Triton X-100) and/or enzymes (e.g. lysozyme).^[2] Finally, the cell debris can be removed by centrifugation so that the proteins and other soluble compounds remain in the supernatant.

Also proteases are released during cell lysis, which will start digesting the proteins in the solution. If the protein of interest is sensitive to proteolysis, it is recommended to proceed quickly, and to keep the extract cooled, to slow down the digestion. Alternatively, one or more protease inhibitors can be added to the lysis buffer immediately before cell disruption. Sometimes it is also necessary to add DNase in order to reduce the viscosity of the cell lysate caused by a high DNA content.

Precipitation and differential solubilization

In bulk protein purification, a common first step to isolate proteins is precipitation with ammonium sulfate $(\text{NH}_4)_2\text{SO}_4$. This is performed by adding increasing amounts of ammonium sulfate and collecting the different fractions of precipitate protein. Ammonium sulfate can be removed by dialysis. The hydrophobic groups on the proteins get exposed to the atmosphere, attract other protein hydrophobic groups and get aggregated. Protein precipitated will be large enough to be visible. One advantage of this method is that it can be performed inexpensively with very large volumes.

The first proteins to be purified are water-soluble proteins. Purification of integral membrane proteins requires disruption of the cell membrane in order to isolate any one particular protein from others that are in the same membrane compartment. Sometimes a particular membrane fraction can be isolated first, such as isolating mitochondria from cells before purifying a protein located in a mitochondrial membrane. A detergent such as sodium dodecyl sulfate (SDS) can be used to dissolve cell membranes and keep membrane proteins in solution

during purification; however, because SDS causes denaturation, milder detergents such as Triton X-100 or CHAPS can be used to retain the protein's native conformation during complete purification.

Ultracentrifugation

Centrifugation is a process that uses centripetal force to separate mixtures of particles of varying masses or densities suspended in a liquid. When a vessel (typically a tube or bottle) containing a mixture of proteins or other particulate matter, such as bacterial cells, is rotated at high speeds, the inertia of each particle yields an force in the direction of the particles velocity that is proportional to its mass. The tendency of a given particle to move through the liquid because of this force is offset by the resistance the liquid exerts on the particle. The net effect of "spinning" the sample in a centrifuge is that massive, small, and dense particles move outward faster than less massive particles or particles with more "drag" in the liquid. When suspensions of particles are "spun" in a centrifuge, a "pellet" may form at the bottom of the vessel that is enriched for the most massive particles with low drag in the liquid.

Non-compacted particles remain mostly in the liquid called "supernatant" and can be removed from the vessel thereby separating the supernatant from the pellet. The rate of centrifugation is determined by the angular acceleration applied to the sample, typically measured in comparison to the g. If samples are centrifuged long enough, the particles in the vessel will reach equilibrium wherein the particles accumulate specifically at a point in the vessel where their buoyant density is balanced with centrifugal force. Such an "equilibrium" centrifugation can allow extensive purification of a given particle.

Sucrose gradient centrifugation — a linear concentration gradient of sugar (typically sucrose, glycerol, or a silica based density gradient media, like Percoll) is generated in a tube such that the highest concentration is on the bottom and lowest on top. Percoll is a trademark owned by GE Healthcare companies. A protein sample is then layered on top of the gradient and spun at high speeds in an ultracentrifuge. This causes heavy macromolecules to migrate towards the bottom of the tube faster than lighter material. During centrifugation in the absence of sucrose, as particles move farther and farther from the center of rotation, they experience more and more centrifugal force (the further they move, the faster they move). The problem with this is that the useful separation range of within the vessel is restricted to a small observable window. Spinning a sample twice as long doesn't mean the particle of interest will go twice as far, in fact,

Protein purification and separation

it will go significantly further. However, when the proteins are moving through a sucrose gradient, they encounter liquid of increasing density and viscosity. A properly designed sucrose gradient will counteract the increasing centrifugal force so the particles move in close proportion to the time they have been in the centrifugal field. Samples separated by these gradients are referred to as "rate zonal" centrifugations. After separating the protein/particles, the gradient is then fractionated and collected.

Purification strategies



Chromatographic equipment. Here set up for a size exclusion chromatography. The buffer is pumped through the column (right) by a computer controlled device.

Choice of a starting material is key to the design of a purification process. In a plant or animal, a particular protein usually isn't distributed homogeneously throughout the body; different organs or tissues have higher or lower concentrations of the protein. Use of only the tissues or organs with the highest concentration decreases the volumes needed to produce a given amount of purified protein. If the protein is present in low abundance, or if it has a high value, scientists may use recombinant DNA technology to develop cells that will produce large quantities of the desired protein (this is known as an expression system). Recombinant expression allows the protein to be tagged, e.g. by a His-tag, to facilitate purification, which means that the purification can be done in fewer steps. In addition, recombinant expression usually starts with a higher fraction of the desired protein than is present in a natural source.

An analytical purification generally utilizes three properties to separate proteins. First, proteins may be purified according to their isoelectric points by running them through a pH

graded gel or an ion exchange column. Second, proteins can be separated according to their size or molecular weight via size exclusion chromatography or by SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) analysis. Proteins are often purified by using 2D-PAGE and are then analysed by peptide mass fingerprinting to establish the protein identity. This is very useful for scientific purposes and the detection limits for protein are nowadays very low and nanogram amounts of protein are sufficient for their analysis. Thirdly, proteins may be separated by polarity/hydrophobicity via high performance liquid chromatography or reversed-phase chromatography.

Usually a protein purification protocol contains one or more chromatographic steps. The basic procedure in chromatography is to flow the solution containing the protein through a column packed with various materials. Different proteins interact differently with the column material, and can thus be separated by the time required to pass the column, or the conditions required to elute the protein from the column. Usually proteins are detected as they are coming off the column by their absorbance at 280 nm. Many different chromatographic methods exist:

Size exclusion chromatography

Chromatography can be used to separate protein in solution or denaturing conditions by using porous gels. This technique is known as size exclusion chromatography. The principle is that smaller molecules have to traverse a larger volume in a porous matrix. Consequentially, proteins of a certain range in size will require a variable volume of eluent (solvent) before being collected at the other end of the column of gel.

In the context of protein purification, the eluent is usually pooled in different test tubes. All test tubes containing no measurable trace of the protein to purify are discarded. The remaining solution is thus made of the protein to purify and any other similarly-sized proteins.

Separation based on charge or hydrophobicity

Hydrophobic interaction chromatography

HIC media is amphiphilic, with both hydrophobic and hydrophilic regions, allowing for separation of proteins based on their surface hydrophobicity. In pure water, the interactions between the resin and the hydrophobic regions of protein would be very weak, but this interaction is enhanced by applying a protein sample to HIC resin in high ionic strength buffer. The ionic strength of the buffer is then reduced to elute proteins in order of decreasing hydrophobicity.^[3]

Ion exchange chromatography

Ion exchange chromatography separates compounds according to the nature and degree of their ionic charge. The column to be used is selected according to its type and strength of charge. Anion exchange resins have a positive charge and are used to retain and separate negatively charged compounds (anions), while cation exchange resins have a negative charge and are used to separate positively charged molecules (cations).

Before the separation begins a buffer is pumped through the column to equilibrate the opposing charged ions. Upon injection of the sample, solute molecules will exchange with the buffer ions as each competes for the binding sites on the resin. The length of retention for each solute depends upon the strength of its charge. The most weakly charged compounds will elute first, followed by those with successively stronger charges. Because of the nature of the separating mechanism, pH, buffer type, buffer concentration, and temperature all play important roles in controlling the separation.

Ion exchange chromatography is a very powerful tool for use in protein purification and is frequently used in both analytical and preparative separations.



Nickel-affinity column. The resin is blue since it has bound nickel.

Free-flow-electrophoresis

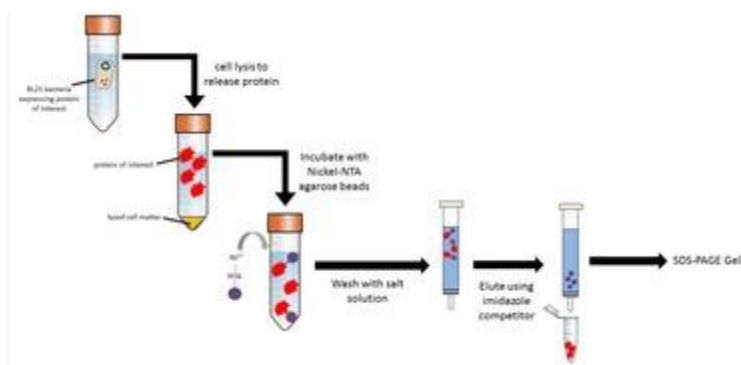
Free-flow electrophoresis (FFE) is a carrier-free electrophoresis technique that allows preparative protein separation in a laminar buffer stream by using an orthogonal electric field. By making use of a pH-gradient, that can for example be induced by ampholytes, this technique allows to separate protein isoforms up to a resolution of < 0.02 delta-pI.

Affinity chromatography

Affinity Chromatography is a separation technique based upon molecular conformation, which frequently utilizes application specific resins. These resins have ligands attached to their surfaces which are specific for the compounds to be separated. Most frequently, these ligands function in a fashion similar to that of antibody-antigen interactions. This "lock and key" fit between the ligand and its target compound makes it highly specific, frequently generating a single peak, while all else in the sample is unretained.

Many membrane proteins are glycoproteins and can be purified by lectin affinity chromatography. Detergent-solubilized proteins can be allowed to bind to a chromatography resin that has been modified to have a covalently attached lectin. Proteins that do not bind to the lectin are washed away and then specifically bound glycoproteins can be eluted by adding a high concentration of a sugar that competes with the bound glycoproteins at the lectin binding site. Some lectins have high affinity binding to oligosaccharides of glycoproteins that is hard to compete with sugars, and bound glycoproteins need to be released by denaturing the lectin.

Metal binding



Schematic showing the steps involved in a metal binding strategy for protein purification. The use of nickel immobilized with Nitrilotriacetic acid (NTA) is shown here.

A common technique involves engineering a sequence of 6 to 8 histidines into the N- or C-terminal of the protein. The polyhistidine binds strongly to divalent metal ions such as nickel and cobalt. The protein can be passed through a column containing immobilized nickel ions, which binds the polyhistidine tag. All untagged proteins pass through the column. The protein can be eluted with imidazole, which competes with the polyhistidine tag for binding to the column, or by a decrease in pH (typically to 4.5), which decreases the affinity of the tag for

the resin. While this procedure is generally used for the purification of recombinant proteins with an engineered affinity tag (such as a 6xHis tag or Clontech's HAT tag), it can also be used for natural proteins with an inherent affinity for divalent cations.

Immunoaffinity chromatography



A HPLC. From left to right: A pumping device generating a gradient of two different solvents, a steel enforced column and an apparatus for measuring the absorbance.

Immunoaffinity chromatography uses the specific binding of an antibody-antigen to selectively purify the target protein. The procedure involves immobilizing a protein to a solid substrate (e.g. a porous bead or a membrane), which then selectively binds the target, while everything else flows through. The target protein can be eluted by changing the pH or the salinity. The immobilized ligand can be an antibody (such as Immunoglobulin G) or it can be a protein (such as Protein A). Because this method does not involve engineering in a tag, it can be used for proteins from natural sources.^[4]

Purification of a tagged protein

Another way to tag proteins is to engineer an antigen peptide tag onto the protein, and then purify the protein on a column or by incubating with a loose resin that is coated with an immobilized antibody. This particular procedure is known as immunoprecipitation. Immunoprecipitation is quite capable of generating an extremely specific interaction which usually results in binding only the desired protein. The purified tagged proteins can then easily be separated from the other proteins in solution and later eluted back into clean solution.

When the tags are not needed anymore, they can be cleaved off by a protease. This often involves engineering a protease cleavage site between the tag and the protein.

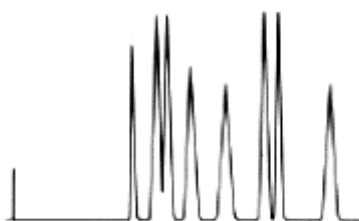
High Performance Liquid Chromatography

What is Liquid Chromatography?

The history of Liquid Chromatography (LC) began at early 20th century. In 1906, a Russian botanist Mikhail Tswett invented LC in order to separate various plant pigments. He injected the plant extract and petroleum ether through a glass column packed with calcium carbonate. Since this was done on a glass column, he was able to observe the changes inside the column. At the beginning, there is only one layer of pigment on the top of the column (Figure 1.a). But as time passes by, the pigment is separated into four different colored-layers (Figure 1.b). The later research discovered that those four-layers were (i) bluish green; chlorophyll a, (ii) yellowish green; chlorophyll b, (iii) yellow; xanthine, and (iv) orange; carotene. The whole process of separation took several hours and thus it was not a very practical method. This long analysis time was a part of the reason that LC did not become a popular analytical tool until 1970s, a half century after Mikhail Tswett's invention. Below is some terms commonly used in chromatography analysis. Chromatography is a separation technique and the word chromatography originated from chroma meaning "color" and graphein meaning "write". Chromatograph is the separation equipment and chromatogram is an out-put chart obtained from the analysis.



Chromatograph



chromatogram

Figure 2. Images of chromatography, chromatograph, and chromatogram

What is HPLC?

HPLC stands for High Performance Liquid Chromatography. Before HPLC was available, LC analysis was carried by gravitational flow of the eluent (the solvent used for LC analysis) thus required several hours for the analysis to be completed. Even the improvements added in later time were able to shorten the analysis time slightly. Those classical/initial LC systems are called "low pressure chromatography" or "column chromatography".

In 1970s in the US, Jim Waters founded Waters Corporation and started to sell HPLC instruments. This promoted the use of HPLC in practical analysis areas. The LC systems that Waters Corporation developed used high-pressure pump that generates rapid-flow of eluent, and thus resulted in dramatic

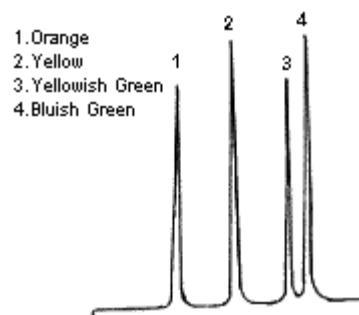


Figure 3. Representation of Tswett's LC analysis.

improvement in the analysis time. Compared to the "low pressure chromatography" the newer types were called "high pressure liquid chromatography". Therefore it was used to be thought that HPLC stands for High Pressure Liquid Chromatography, however nowadays it is a common agreement that HPLC stands for High Performance Liquid Chromatography. Another big change from Tswett's date was the data acquisition methods. Instead of observing the changes of layers by eyes, detector system was coupled to the LC and out-put was recorded on paper chart. If we were to demonstrate Tswett's analysis result on a chart (chromatogram), it will be like figure 3.

Initially, HPLC system was referred to Waters Corporation's system. Still now, Waters Corporation is the HPLC pioneer, but there are several other companies that manufacture and sell HPLC systems.

Technically speaking, the word LC represents all the Liquid Chromatography, including low pressure LC, however most LC systems used these days are HPLC thus often the word LC is used as comparable as HPLC.

Components of HPLC

Typical HPLC system consists of followings (Figure 4). Details of each are explained below.

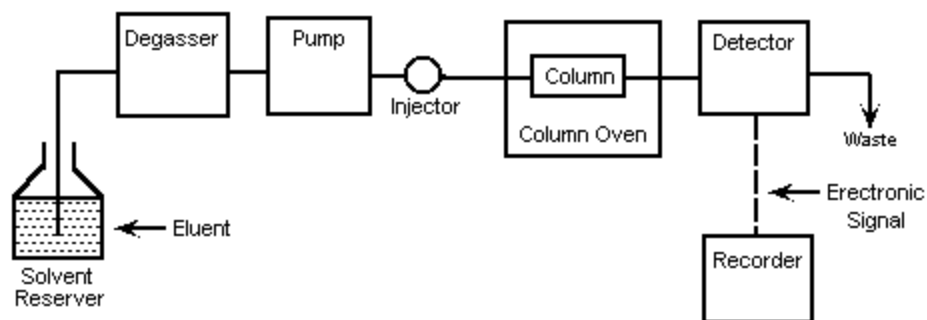


Figure 4. Components of HPLC system

Pump

In the earlier state of HPLC development, the pump was the most important part of the system. The development of HPLC can be said that it was a development of pump system. Pump is positioned in the most upper stream of the LC system and generates a flow of eluent from the solvent reservoir to the system. In the earlier stage of LC development, to be able to generate the high pressure was the one of the most important system requirements. However, nowadays, the high pressure generation is a "standard" requirement and what is more concerned nowadays is to be able to provide a consistent pressure at any condition, to provide a controllable and reproducible flow rate. Since a change in the flow rate can influence the analysis largely. Most pumps used in current LC systems generate the flow by back-and-forth motion of a motor-driven piston (reciprocating pumps). Because of this piston motion, it produces "pulses". There have been large system improvements to reduce this pulsation and the recent pumps create much less pulse compared to the older ones. However, recent analysis requires very high sensitivity to quantify a small amount of analytes, and thus even a minor change in the flow rate can influence the analysis. Therefore, the pumps required for the high sensitivity analysis needs to be highly precise.

Injector

An injector is placed next to the pump. The simplest method is to use a syringe, and the sample is introduced to the flow of eluent. Since the precision of LC measurement is largely affected by the reproducibility of sample injection, the design of injector is an important factor. The most

widely used injection method is based on sampling loops. The use of autosampler (auto-injector) system is also widely used that allows repeated injections in a set scheduled-timing.

Column

The separation is performed inside the column; therefore, it can be said that the column is the heart of an LC system. The theory of chromatography column has not changed since Tswett's time; however there has been continuous improvement in column development. The recent columns are often prepared in stainless steel housing, instead of glass columns used in Tswett's experiment. The packing material generally used is silica or polymer gels compared to calcium carbonate used by Tswett. The eluent used for LC varies from acidic to basic solvents. Most column housing is made of stainless steel, since stainless is tolerant towards a large variety of solvents. However, for the analysis of some analytes such as biomolecules and ionic compounds, contact with metal is not desired, thus a polyether ether ketone (PEEK) column housing is used instead.

Detector

Separation of analytes is performed inside the column, whereas a detector is used to observe the obtained separation. The composition of the eluent is consistent when no analyte is present. While the presence of analyte changes the composition of the eluent. What detector does is to measure these differences. This difference is monitored as a form of electronic signal. There are different types of detectors available. Different detector types are explained in Lesson 6.

Recorder

The change in eluent detected by a detector is in the form of electronic signal, and thus it is still not visible to our eyes. In older days, pen (paper)-chart recorder was popularly used. Nowadays, computer based data processor (integrator) is more common. There are various types of data processors; examples include a simple system consisting of in-built printer and word processor, and a personal computer type consisting of display monitor, keyboard, and printer. Also there are

software that are specifically designed for LC system. It provides not only data acquisition, but features like peak-fitting, base line correction, automatic concentration calculation, molecular weight determination, etc...

The components introduced so far are the basics of LC system. Below are some optional equipment used with the basic LC system.

Degasser

The eluent used for LC analysis may contain gases such as oxygen that are non-visible to our eyes. When gas is present in the eluent, this is detected as a noise and causes unstable baseline. Generally used method includes sparging (bubbling of inert gas), use of aspirator, distillation system, and/or heating and stirring. However, the method is not convenient and also when the solvent is left for a certain time period (e.g., during the long analysis), gas will dissolve back gradually. Degasser uses special polymer membrane tubing to remove gases. The numerous very small pores on the surface of the polymer tube allow the air to go through while preventing any liquid to go through the pore. By placing this tubing under low pressure container, it created pressure differences inside and outside the tubing (higher inside the tubing). This difference let the dissolved gas to move through the pores and remove the gas. Compared to classical batch type degassing, the degasser can be used on-line, it is more convenient and efficient. Many of new HPLC unit system contain a degasser.

Column Heater

The LC separation is often largely influenced by the column temperature. In order to obtain repeatable results, it is important to keep the consistent temperature conditions. Also for some analysis, such as sugar and organic acid, better resolutions can be obtained at elevated temperature (50 to 80°C). It is also important to keep stable temperature to obtain repeatable results even it is analyzed at around room temperature. There are possibilities that small different of temperature causes different separation results. Thus columns are generally kept inside the column oven (column heater).

Other Chromatography

Another member of chromatography, as often used as LC, is the Gas Chromatography (GC). While LC uses eluent (flow generated by a pump system), GC uses gas (carrier gas) provided from a gas-tank, thus it does not require pump system. Therefore, GC system is simpler than LC system and for that reason; it was widely available before LC system become popular. At that time, the word chromatography was referred to GC. GC's simple system was an advantage but also a disadvantage. The samples to be analyzed need to be in the gaseous form before introduced to the flow of carrier gas. Thus, if the original sample was in gaseous form, GC is a convenient tool. However most sample is in the form of liquid or solid and requires heating at high temperature to make it into a gas. Some analytes are alternated by heat, so the GC method is not an ideal for that case. Related problem is that GC has an upper limit for the analysis of large molecular weight compounds.

In comparison, liquid samples can be analyzed directly by LC. Also if solid sample can be dissolved in a solvent, it can be analyzed by LC. Some samples are insoluble in water, but most of those are soluble in organic solvent, there is a high chance that we can find a solvent that dissolves the sample. Therefore, LC can be used at ambient temperature (i.e., without causing alternation of analyte) for a wide range of analysis. This made the LC method become more popular than GC method. There are other types of chromatography such as thin layer chromatography, super-critical fluid chromatography, paper chromatography etc... but their use is even less popular than GC.

Mass spectrometry

Mass spectrometry (MS) is an analytical technique that ionizes **chemical species** and sorts the **ions** based on their mass to charge ratio. In simpler terms, a **mass spectrum** measures the masses within a sample. Mass spectrometry is used in many different fields and is applied to pure samples as well as complex mixtures.

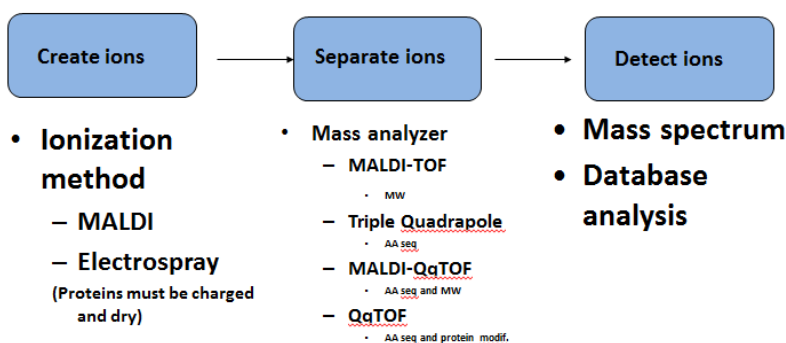
A **mass spectrum** is a plot of the ion signal as a function of the mass-to-charge ratio. These spectra are used to determine the elemental or **isotopic signature** of a sample, the masses of particles and of **molecules**, and to elucidate the chemical structures of molecules, such as **peptides** and other **chemical compounds**.

In a typical MS procedure, a sample, which may be solid, liquid, or gas, is ionized, for example by bombarding it with electrons. This may cause some of the sample's molecules to break into charged fragments. These ions are then separated according to their mass-to-charge ratio, typically by accelerating them and subjecting them to an electric or magnetic field: ions of the same mass-to-charge ratio will undergo the same amount of deflection. The ions are detected by a mechanism capable of detecting charged particles, such as an **electron multiplier**. Results are displayed as spectra of the relative abundance of detected ions as a function of the mass-to-charge ratio. The atoms or molecules in the sample can be identified by correlating known masses to the identified masses or through a characteristic fragmentation pattern.

Parts of a mass spectrometer

A mass spectrometer consists of three components: an ion source, a mass analyzer, and a detector. The **ionizer** converts a portion of the sample into ions. There is a wide variety of ionization techniques, depending on the phase (solid, liquid, gas) of the sample and the efficiency of various ionization mechanisms for the unknown species. An extraction system removes ions from the sample, which are then targeted through the mass analyzer and onto the *detector*. The differences in masses of the fragments allows the mass analyzer to sort the ions by their mass-to-charge ratio. The detector measures the value of an indicator quantity and thus provides data for calculating the abundances of each ion present. Some detectors also give spatial information, e.g., a multichannel plate.

How does a mass spectrometer work?



Creating ions

The **ion source** is the part of the mass spectrometer that ionizes the material under analysis (the analyte). The ions are then transported by **magnetic** or **electric fields** to the mass analyzer.

Techniques for ionization have been key to determining what types of samples can be analyzed by mass spectrometry. **Electron ionization** and **chemical ionization** are used for **gases** and **vapors**. In chemical ionization sources, the analyte is ionized by chemical ion-molecule reactions during collisions in the source. Two techniques often used with **liquid** and **solid** biological samples include **electrospray ionization** (invented by **John Fenn**) and **matrix-assisted laser desorption/ionization** (MALDI, initially developed as a similar technique "Soft Laser Desorption (SLD)" by K. Tanaka for which a Nobel Prize was awarded and as MALDI by M. Karas and F. Hillenkamp

Hard ionization and soft ionization

In mass spectrometry, ionization refers to the production of gas phase ions suitable for resolution in the mass analyser or mass filter. Ionization occurs in the **ion source**. There are several **ion sources** available; each has advantages and disadvantages for particular applications. For example, **electron ionization** (EI) gives a high degree of fragmentation, yielding highly detailed mass spectra which when skilfully analysed can provide important information for structural elucidation/characterisation and facilitate identification of unknown compounds by comparison to mass spectral libraries obtained under identical operating conditions. However, EI is not suitable for coupling to **HPLC**, i.e. **LC-MS**, since at atmospheric pressure, the filaments used to

generate electrons burn out rapidly. Thus EI is coupled predominantly with [GC](#), i.e. [GC-MS](#), where the entire system is under high vacuum.

Hard ionization techniques are processes which impart high quantities of residual energy in the subject molecule invoking large degrees of fragmentation (i.e. the systematic rupturing of bonds acts to remove the excess energy, restoring stability to the resulting ion). Resultant ions tend to have m/z lower than the molecular mass (other than in the case of proton transfer and not including isotope peaks). The most common example of hard ionization is electron ionization (EI).

Soft ionization refers to the processes which impart little residual energy onto the subject molecule and as such result in little fragmentation. Examples include [fast atom bombardment](#) (FAB), [chemical ionization](#) (CI), [atmospheric-pressure chemical ionization](#) (APCI), [electrospray ionization](#) (ESI), and [matrix-assisted laser desorption/ionization](#) (MALDI)

Inductively coupled plasma

[Inductively coupled plasma](#) (ICP) sources are used primarily for cation analysis of a wide array of sample types. In this source, a plasma that is electrically neutral overall, but that has had a substantial fraction of its atoms ionized by high temperature, is used to atomize introduced sample molecules and to further strip the outer electrons from those atoms. The plasma is usually generated from argon gas, since the first ionization energy of argon atoms is higher than the first of any other elements except He, O, F and Ne, but lower than the second ionization energy of all except the most electropositive metals. The heating is achieved by a radio-frequency current passed through a coil surrounding the plasma.

Other ionization techniques

Others include [photoionization](#), [glow discharge](#), [field desorption](#) (FD), [fast atom bombardment](#) (FAB), [thermospray,desorption/ionization on silicon](#) (DIOS), [Direct Analysis in Real Time](#) (DART), [atmospheric pressure chemical ionization](#) (APCI), [secondary ion mass spectrometry](#) (SIMS), [spark ionization](#) and [thermal ionization](#) (TIMS)

Matrix-assisted laser desorption/ionization (MALDI)

Matrix-assisted laser desorption/ionization (MALDI) is a soft ionization technique used in mass spectrometry, allowing the analysis of biomolecules (biopolymers such as DNA, proteins, peptides and sugars) and large organic molecules (such as polymers, dendrimers and other macromolecules), which tend to be fragile and fragment when ionized by more conventional ionization methods. It is similar in character to electrospray ionization (ESI) in that both techniques are relatively soft ways of obtaining ions of large molecules in the gas phase, though MALDI produces far fewer multiply charged ions.

MALDI methodology is a three-step process. First, the sample is mixed with a suitable matrix material and applied to a metal plate. Second, a pulsed laser irradiates the sample, triggering ablation and desorption of the sample and matrix material. Finally, the analyte molecules are ionized by being protonated or deprotonated in the hot plume of ablated gases, and can then be accelerated into whichever mass spectrometer is used to analyse them

The term matrix-assisted laser desorption ionization (MALDI) was coined in 1985 by [Franz Hillenkamp](#), [Michael Karas](#) and their colleagues.^[2] These researchers found that the [amino acid alanine](#) could be ionized more easily if it was mixed with the amino acid [tryptophan](#) and irradiated with a pulsed 266 nm laser. The tryptophan was absorbing the laser energy and helping to ionize the non-absorbing alanine. Peptides up to the 2843 Da peptide [melittin](#) could be ionized when mixed with this kind of “matrix”.^[3] The breakthrough for large molecule laser desorption ionization came in 1987 when [Koichi Tanaka](#) of Shimadzu Corporation and his co-workers used what they called the “ultra fine metal plus liquid matrix method” that combined 30 nm [cobalt](#) particles in [glycerol](#) with a 337 nm [nitrogen laser](#) for ionization.^[4] Using this laser and matrix combination, Tanaka was able to ionize biomolecules as large as the 34,472 Da protein carboxypeptidase-A. Tanaka received one-quarter of the 2002 [Nobel Prize in Chemistry](#) for demonstrating that, with the proper combination of laser wavelength and matrix, a protein can be ionized.^[5] Karas and Hillenkamp were subsequently able to ionize the 67 kDa protein albumin using a nicotinic acid matrix and a 266 nm laser.^[6] Further improvements were realized through the use of a 355 nm laser and the [cinnamic acid](#) derivatives [ferulic acid](#), [caffeic acid](#) and [sinapinic acid](#) as the matrix.^[7] The availability of small and relatively inexpensive nitrogen lasers operating at 337 nm wavelength and the first

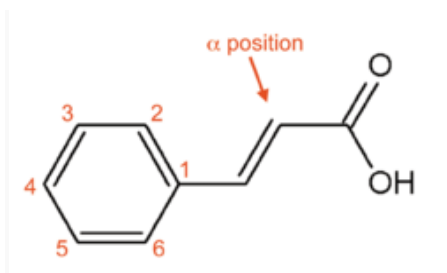
commercial instruments introduced in the early 1990s brought MALDI to an increasing number of researchers.^[8] Today, mostly organic matrices are used for MALDI mass spectrometry.

Matrix and sample preparation

UV MALDI Matrix List				
Compound	Other Names	Solvent	Wavelength (nm)	Applications
<u>2,5-dihydroxybenzoic acid</u> ^[9]	DHB, <u>Gentisic acid</u>	<u>acetonitrile</u> , <u>water</u> , <u>methanol</u> , <u>acetone</u> , <u>chloroform</u>	337, 355, 266	<u>peptides</u> , <u>nucleotides</u> , <u>oligonucleotides</u> , <u>oligosaccharides</u>
<u>3,5-dimethoxy-4-hydroxycinnamic acid</u> ^{[7][10]}	sinapic acid; <u>sinapinic acid</u> ; SA	<u>acetonitrile</u> , <u>water</u> , <u>acetone</u> , <u>chloroform</u>	337, 355, 266	peptides, proteins, <u>lipids</u>
<u>4-hydroxy-3-methoxycinnamic acid</u> ^{[7][10]}	<u>ferulic acid</u>	<u>acetonitrile</u> , <u>water</u> , <u>propanol</u>	337, 355, 266	proteins
<u>α-Cyano-4-hydroxycinnamic acid</u> ^[11]	CHCA	<u>acetonitrile</u> , <u>water</u> , <u>ethanol</u> , <u>acetone</u>	337, 355	peptides, lipids, nucleotides
<u>Picolinic acid</u> ^[12]	PA	Ethanol	266	oligonucleotides
<u>3-hydroxy</u>	HPA	Ethanol	337,	oligonucleotides

<u>picolinic acid</u> ^[13]			355	
---	--	--	-----	--

The matrix consists of [crystallized](#) molecules, of which the three most commonly used are 3,5-dimethoxy-4-hydroxycinnamic [acid](#) ([sinapinic acid](#)), [\$\alpha\$ -cyano-4-hydroxycinnamic acid](#) (CHCA, alpha-cyano or alpha-matrix) and [2,5-dihydroxybenzoic acid](#) (DHB). A [solution](#) of one of these molecules is made, often in a mixture of highly purified [water](#) and an organic [solvent](#) such as [acetonitrile](#) (ACN) or [ethanol](#). A counter ion source such as [Trifluoroacetic acid](#) (TFA) is usually added to generate the [M+H] ions. A good example of a matrix-solution would be 20 [mg/mL](#) sinapinic acid in ACN:water:TFA (50:50:0.1).



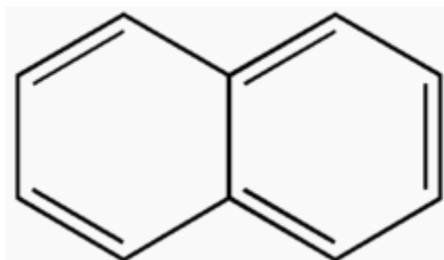
Notation for cinnamic acid substitutions.

The identification of suitable matrix compounds is determined to some extent by trial and error, but they are based on some specific molecular design considerations:

- They are of a fairly low molecular weight (to allow easy vaporization), but are large enough (with a low enough vapor pressure) not to evaporate during sample preparation or while standing in the spectrometer.
- They are often acidic, therefore act as a proton source to encourage ionization of the analyte. Basic matrices have also been reported.^[15]
- They have a strong optical absorption in either the UV or IR range,^[16] so that they rapidly and efficiently absorb the laser irradiation. This efficiency is commonly associated with chemical structures incorporating several [conjugated double bonds](#), as seen in the structure of cinnamic acid.
- They are functionalized with polar groups, allowing their use in aqueous solutions.
- They typically contain a chromophore.

The matrix solution is mixed with the analyte (e.g. [protein](#)-sample). A mixture of water and organic solvent allows both [hydrophobic](#) and water-soluble ([hydrophilic](#)) molecules to dissolve into the solution. This solution is spotted onto a MALDI plate (usually a metal plate

designed for this purpose). The solvents vaporize, leaving only the recrystallized matrix, but now with analyte molecules embedded into MALDI crystals. The matrix and the analyte are said to be co-crystallized. Co-crystallization is a key issue in selecting a proper matrix to obtain a good quality mass spectrum of the analyte of interest.



Naphthalene and naphthalene-like compounds can also be used as a matrix to ionize a sample

The matrix can be used to tune the instrument to ionize the sample in different ways. As mentioned above, acid-base like reactions are often utilized to ionize the sample, however, molecules with conjugated π systems, such as naphthalene like compounds, can also serve as an electron acceptor and thus a matrix for MALDI/TOF. This is particularly useful in studying molecules that also possess [conjugated \$\pi\$ systems](#). The most widely used application for these matrices is studying porphyrin like compounds such as [chlorophyll](#). These matrices have been shown to have better ionization patterns that do not result in odd fragmentation patterns or complete loss of side chains. It has also been suggested that conjugated porphyrin like molecules can serve as a matrix and cleave themselves eliminating the need for a separate matrix compound

Technology and instrumentation

There are several variations of the MALDI technology and comparable instruments are today produced for very different purposes. From more academic and analytical, to more industrial and high throughput. The MS field has expanded into requiring ultrahigh resolution mass spectrometry such as the FT-ICR instruments^{[21][22]} as well as more high-throughput instruments.^[23] As many MALDI MS instruments can be bought with an interchangeable ionization source ([Electrospray ionization](#), MALDI, [Atmospheric pressure ionization](#), etc) the technologies often overlap and many times any soft ionization method could potentially be used. For more variations of soft ionization methods go to [Soft laser desorption](#) or [Ion source](#).

Laser

MALDI techniques typically employ the use of UV [lasers](#) such as nitrogen lasers (337 nm) and frequency-tripled and quadrupled [Nd:YAG lasers](#) (355 nm and 266 nm respectively). Although not as common, infrared lasers are used due to their softer mode of ionization. IR-MALDI also has the advantage of greater material removal (useful for biological samples), less low-mass interferences, and compatibility with other matrix-free laser desorption mass spectrometry methods.^[24]

Ionization mechanism

The [laser](#) is fired at the matrix crystals in the dried-droplet spot. The matrix absorbs the laser energy and it is thought that primarily the matrix is desorbed and ionized (by addition of a [proton](#)) by this event. The hot plume produced during ablation contains many species: neutral and ionized matrix molecules, protonated and deprotonated matrix molecules, matrix clusters and [nanodroplets](#). Ablated species may participate in the ionization of analyte, though the mechanism of MALDI is still debated. The matrix is then thought to transfer protons to the analyte molecules (e.g., protein molecules), thus charging the analyte.^[25] An ion observed after this process will consist of the initial neutral molecule [M] with ions added or removed. This is called a quasimolecular ion, for example $[M+H]^+$ in the case of an added proton, $[M+Na]^+$ in the case of an added [sodium](#) ion, or $[M-H]^-$ in the case of a removed proton. MALDI is capable of creating singly charged ions or multiply charged ions ($[M+nH]^{n+}$) depending on the nature of the matrix, the laser intensity, and/or the voltage used. Note that these are all even-electron species. Ion signals of radical cations (photoionized molecules) can be observed, e.g., in the case of matrix molecules and other organic molecules.

A recent method termed matrix-assisted ionization [MAI] uses matrix preparation identical to MALDI but does not require laser ablation to produce analyte ions of volatile or nonvolatile compounds. Simply exposing the matrix [e.g. 3-nitrobenzonitrile] with analyte to the vacuum of the mass spectrometer creates ions with nearly identical charge states to electrospray ionization. It is suggested that there are likely mechanistic commonality between this process and MALDI.

Time of Flight

The type of a mass spectrometer most widely used with MALDI is the [TOF](#) (time-of-flight mass spectrometer), mainly due to its large mass range. The TOF measurement procedure

is also ideally suited to the MALDI ionization process since the pulsed laser takes individual 'shots' rather than working in continuous operation. MALDI-TOF instrument or [reflectron](#) is equipped with an "ion mirror" that reflects ions using an electric field, thereby doubling the ion flight path and increasing the resolution. Today, commercial [reflectron](#) TOF instruments reach a resolving power $m/\Delta m$ of well above 20,000 FWHM (full-width half-maximum, Δm defined as the peak width at 50% of peak height).

MALDI has been coupled with [IMS](#)-TOF MS to identify phosphorylated and non-phosphorylated peptides. MALDI-[FT-ICR](#) MS has been demonstrated to be a useful technique where high resolution MALDI-MS measurements are desired.

Atmospheric pressure matrix-assisted laser desorption/ionization

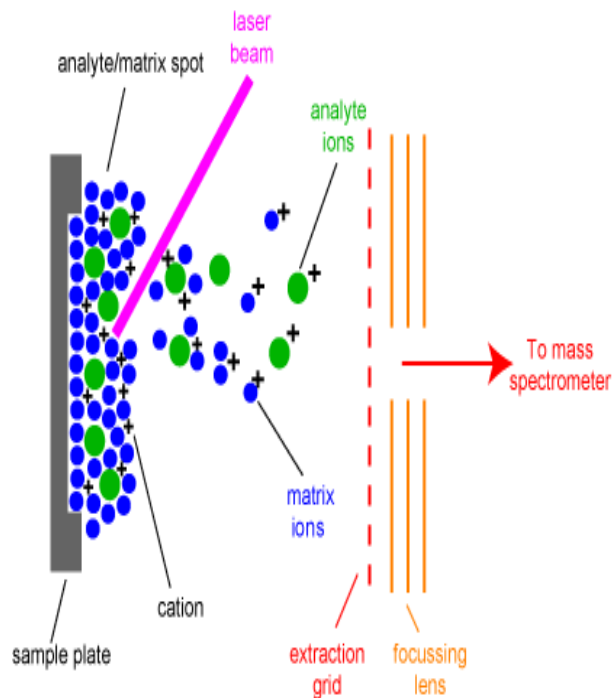
Atmospheric pressure (AP) matrix-assisted laser desorption/ionization (MALDI) is an ionization technique (ion source) that in contrast to vacuum MALDI operates at normal atmospheric environment.^[31] The main difference between vacuum MALDI and AP-MALDI is the pressure in which the ions are created. In vacuum MALDI, ions are typically produced at 10 mTorr or less while in AP-MALDI ions are formed in [atmospheric pressure](#). In the past the main disadvantage of AP MALDI technique compared to the conventional vacuum MALDI has been its limited sensitivity; however, ions can be transferred into the mass spectrometer with high efficiency and attomole detection limits have been reported.^[32] AP-MALDI is used in mass spectrometry (MS) in a variety of applications ranging from proteomics to drug discovery. Popular topics that are addressed by AP-MALDI mass spectrometry include: proteomics; mass analysis of DNA, RNA, PNA, lipids, oligosaccharides, phosphopeptides, bacteria, small molecules and synthetic polymers, similar applications as available also for vacuum MALDI instruments. The AP-MALDI ion source is easily coupled to an ion trap mass spectrometer^[33] or any other MS system equipped with [ESI](#) (electrospray ionization) or nanoESI source

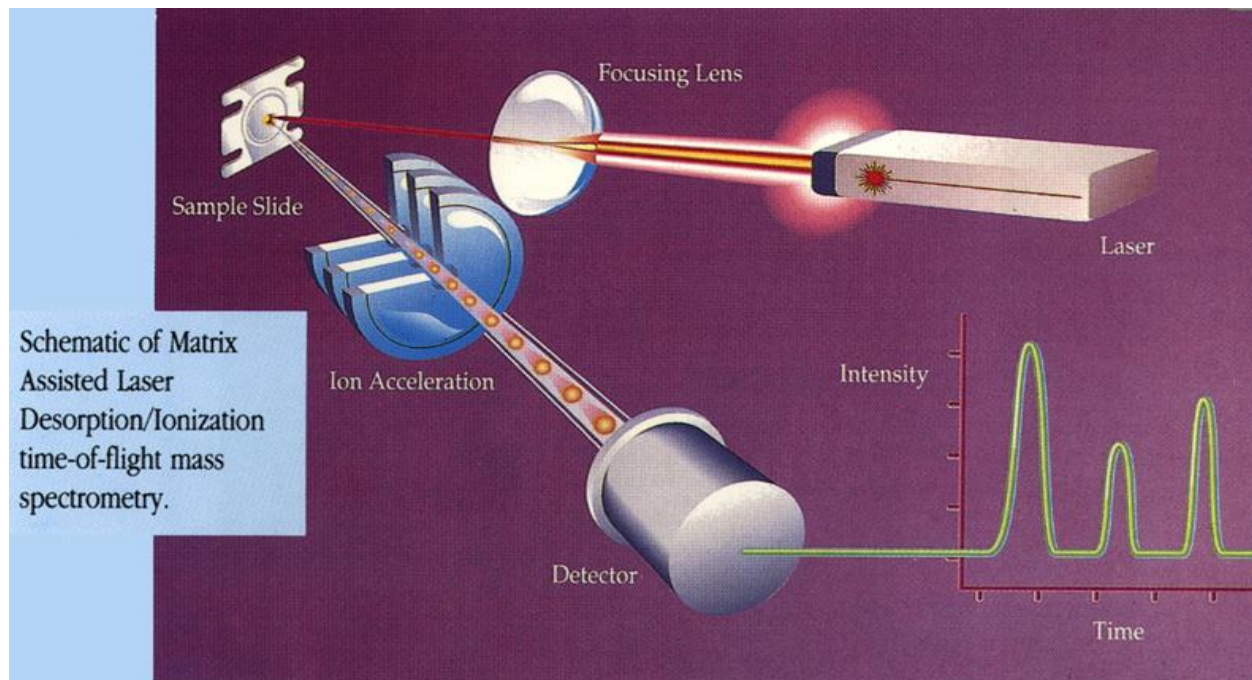
Matrix Assisted Laser Desorption/Ionization (MALDI)

- Method where a laser is used to generate ions of high molecular weight samples, such as proteins and polymers.
- Analyte is embedded in to crystal matrix

Protein purification and separation

- The presence of an aromatic matrix causes the large molecules to ionize instead of decomposing.
- It may involve absorption of light by the matrix
- Transfer of this energy to the analyte
 - which then ionizes into the gas phase as a result of the relatively large amount of energy absorbed.
 - To accelerate the resulting ions into a flight-tube in the mass spectrometer they are subjected to a high electrical field

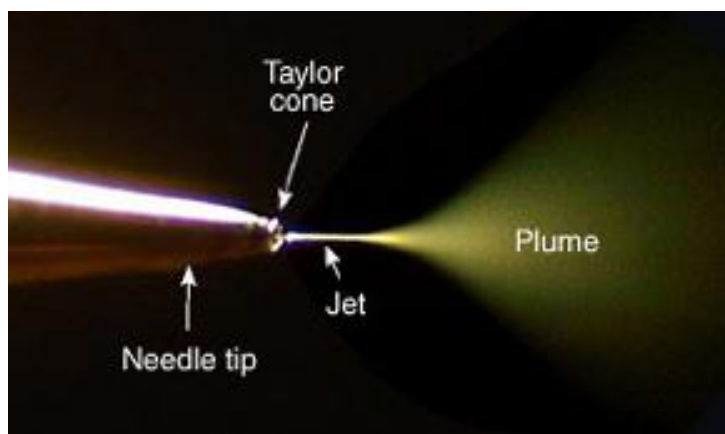




Electrospray ionization

Electrospray ionization (ESI) is a technique used in **mass spectrometry** to produce ions using an **electrospray** in which a high voltage is applied to a liquid to create an **aerosol**. It is especially useful in producing ions from **macromolecules** because it overcomes the propensity of these molecules to fragment when ionized. ESI is different from other atmospheric pressure ionization processes (e.g. **MALDI**) since it may produce multiply charged ions, effectively extending the mass range of the analyser to accommodate the **kDa-MDa** orders of magnitude observed in proteins and their associated polypeptide fragments.

Mass spectrometry using ESI is called electrospray ionization mass spectrometry (ESI-MS) or, less commonly, electrospray mass spectrometry (ES-MS). ESI is a so-called 'soft ionization' technique, since there is very little fragmentation. This can be advantageous in the sense that the molecular ion (or more accurately a pseudo molecular ion) is always observed, however very little structural information can be gained from the simple mass spectrum obtained. This disadvantage can be overcome by coupling ESI with **tandem mass spectrometry** (ESI-MS/MS). Another important advantage of ESI is that solution-phase information can be retained into the gas-phase.



The electrospray ionization technique was first reported by Masamichi Yamashita and John Fenn in 1984. The development of electrospray ionization for the analysis of biological macromolecules was rewarded with the attribution of the **Nobel Prize in Chemistry** to **John Bennett Fenn** in 2002. One of the original instruments used by Dr. Fenn is on display at the Chemical Heritage Foundation in Philadelphia, Pennsylvania

Ionization mechanism

The liquid containing the analyte(s) of interest is dispersed by electrospray, into a fine aerosol. Because the ion formation involves extensive solvent evaporation (also termed desolvation), the typical solvents for electrospray ionization are prepared by mixing water with volatile organic compounds (e.g. methanol acetonitrile). To decrease the initial droplet size, compounds that increase the conductivity (e.g. acetic acid) are customarily added to the solution. These species also act to provide a source of protons to facilitate the ionization process. Large-flow electrosprays can benefit from nebulization a heated inert gas such as nitrogen or carbon dioxide in addition to the high temperature of the ESI source. The aerosol is sampled into the first vacuum stage of a mass spectrometer through a capillary carrying a potential difference of approximately 3000V, which can be heated to aid further solvent evaporation from the charged droplets. The solvent evaporates from a charged droplet until it becomes unstable upon reaching its Rayleigh limit. At this point, the droplet deforms as the electrostatic repulsion of like charges, in an ever-decreasing droplet size, becomes more powerful than the surface tension holding the droplet together. At this point the droplet undergoes Coulomb fission, whereby the original droplet 'explodes' creating many smaller, more stable droplets. The new droplets undergo desolvation and subsequently further Coulomb fissions. During the fission, the droplet loses a small percentage of its mass (1.0–2.3%) along with a relatively large percentage of its charge (10–18%).

There are two major theories that explain the final production of gas-phase ions: the ion evaporation model (IEM) and the charge residue model (CRM). The IEM suggests that as the droplet reaches a certain radius the field strength at the surface of the droplet becomes large enough to assist the field desorption of solvated ions. The CRM suggests that electrospray droplets undergo evaporation and fission cycles, eventually leading progeny droplets that contain on average one analyte ion or less. The gas-phase ions form after the remaining solvent molecules evaporate, leaving the analyte with the charges that the droplet carried.

Coulomb Fission:

Assumes that the increased charge density, due to solvent evaporation, causes large droplets to divide into smaller droplets eventually leading to single ions.

Ion Evaporation:

Assumes the increased charge density that results from solvent evaporation causes Coulombic repulsion to overcome the liquid's surface tension, resulting in a release of ions from droplet surfaces

Parameters Influencing Droplet Size

- The radius (R) of an electrosprayed droplet depends upon fluid density (ρ), flow rate (V_f), and surface tension (γ).

$$R \propto (\rho V_f^2 \gamma)^{1/3}$$

- Thus, higher V_f result in larger initial droplet sizes. Larger droplet sizes lead to lower ionization efficiency because the droplets are not so close in size to the Rayleigh limit

Advantages of ESI

- Soft-ionization technique
- Controllable fragmentation
- Readily coupled to liquid separations
- Produces intact non-covalent complexes
- Multiple-charging of analyte
- Capable of ionizing large molecules (to MDa)

Protein purification and separation

Ionization Method	Typical Analytes	Sample Introduction	Mass Range	Method Highlights
Electron Impact (EI)	Relatively small. Volatile.	GC or liquid or solid probe	To 1000 Daltons	Hard method. Provides structural info
Chemical Ionization (CI)	Relatively small. Volatile.	GC or liquid or solid probe	To 1000 Daltons	Soft method. Molecular ion peak $[M+H]^+$
Electrospray (ESI)	Peptides/proteins. Non-volatile.	Liquid Chromatography	To 200,000 Daltons	Soft method. Ions often multiply charged.
Matrix Assisted Laser Desorption (MALDI)	Peptides/proteins. Non-volatile.	Sample mixed in solid matrix	To 500,000 Daltons	Soft method. Very high mass range.
Fast Atom Bombardment (FAB)	Carbs/peptides. Non-volatile.	Sample mixed in viscous matrix	To 6000 Daltons	Soft method, but harder than ESI or MALDI

Mass Analyzer

Mass analyzers separate the ions according to their **mass-to-charge ratio**. The following two laws govern the dynamics of charged particles in electric and magnetic fields in vacuum:

$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \text{ (Lorentz force law);}$$

$$\mathbf{F} = m\mathbf{a} \text{ (Newton's second law of motion in non-relativistic case, i.e. valid only at ion velocity much lower than the speed of light).}$$

Here \mathbf{F} is the force applied to the ion, m is the mass of the ion, \mathbf{a} is the acceleration, Q is the ion charge, \mathbf{E} is the electric field, and $\mathbf{v} \times \mathbf{B}$ is the **vector cross product** of the ion velocity and the magnetic field

Equating the above expressions for the force applied to the ion yields:

$$(m/Q)\mathbf{a} = \mathbf{E} + \mathbf{v} \times \mathbf{B}.$$

This **differential equation** is the classic equation of motion for **charged particles**. Together with the particle's initial conditions, it completely determines the particle's motion in space and time in terms of m/Q . Thus mass spectrometers could be thought of as "mass-to-charge spectrometers". When presenting data, it is common to use the (officially)**dimensionless** m/z , where z is the number of **elementary charges** (e) on the ion ($z=Q/e$). This quantity, although it is informally called the mass-to-charge ratio, more accurately speaking represents the ratio of the mass number and the charge number, z .

There are many types of mass analyzers, using either static or dynamic fields, and magnetic or electric fields, but all operate according to the above differential equation. Each analyzer type has its strengths and weaknesses. Many mass spectrometers use two or more mass analyzers for **tandem mass spectrometry (MS/MS)**. In addition to the more common mass analyzers listed below, there are others designed for special situations.

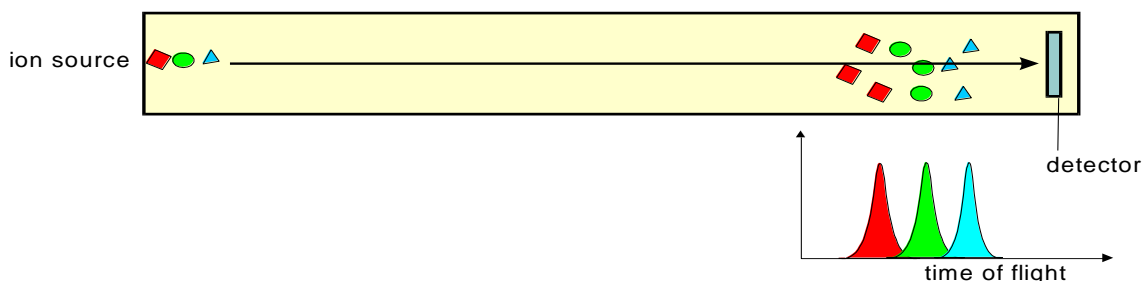
There are several important analyser characteristics. The **mass resolving power** is the measure of the ability to distinguish two peaks of slightly different m/z . The mass accuracy is the ratio of the m/z measurement error to the true m/z . Mass accuracy is usually measured in **ppm** or **milli mass units**. The mass range is the range of m/z amenable to analysis by a given analyzer. The linear dynamic range is the range over which ion signal is linear with analyte concentration. Speed refers to the time frame of the experiment and ultimately is used to determine the number of spectra per unit time that can be generated.

Time-of-flight mass spectrometry

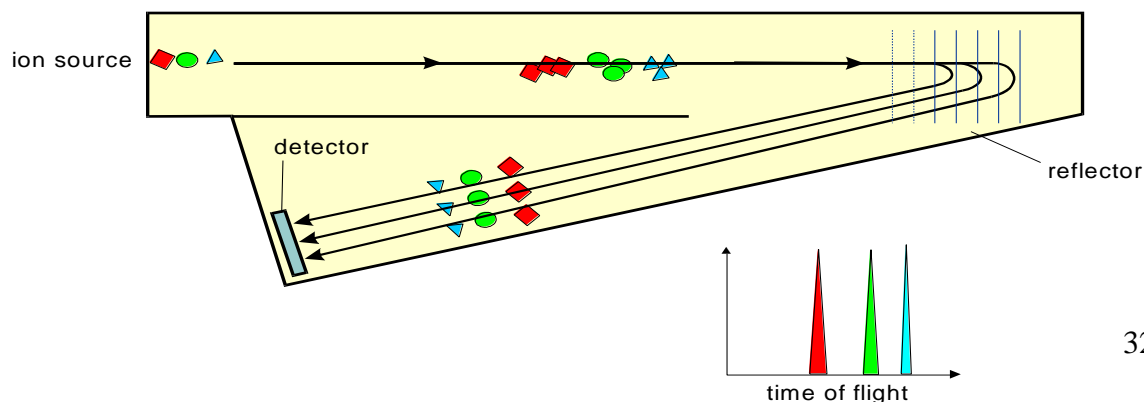
TOF-MS is a method of mass spectrometry in which an ion's mass-to-charge ratio is determined via a time measurement. Ions are accelerated by an electric field of known strength.[1] This acceleration results in an ion having the same kinetic energy as any other ion that has the same charge. The velocity of the ion depends on the mass-to-charge ratio. The time that it subsequently takes for the particle to reach a detector at a known distance is measured. This time will depend on the mass-to-charge ratio of the particle (heavier particles reach lower speeds). From this time and the known experimental parameters one can find the mass-to-charge ratio of the ion.

- The typical detector used with MALDI is the time of flight mass detector (TOF-MS)
 - TOF is a method where the ions are accelerated by an electric field, resulting in ions of the same strength to have the same kinetic energy [7]
 - The time it takes for each ion to traverse the flight tube and arrive at the detector is based on its mass-to-charge ratio; therefore the heavier ions have shorter arrival times compared to lighter ions
- The TOF detector is also equipped with a reflectron, or an ion mirror
 - The reflectron deflects the ion using an electric field and increases the path length, improving signal resolution

Linear Time Of Flight tube

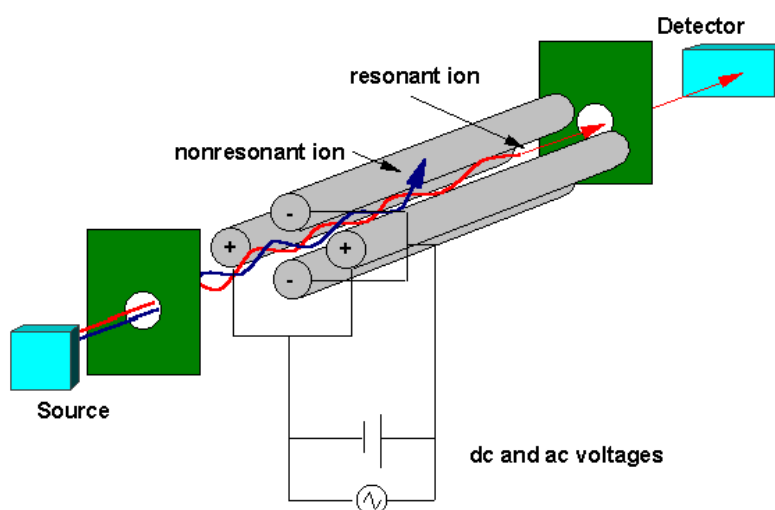


Reflector Time Of Flight tube



Quadrupole Mass Analyzer

The **quadrupole mass analyzer (QMS)** is one type of mass analyzer used in [mass spectrometry](#). It is also known as a **transmission quadrupole mass spectrometer**, **quadrupole mass filter**, or **quadrupole mass spectrometer**. As the name implies, it consists of four cylindrical rods, set parallel to each other. In a quadrupole [mass spectrometer](#) the [quadrupole](#) is the component of the instrument responsible for filtering sample ions, based on their [mass-to-charge ratio](#) (m/z). Ions are separated in a quadrupole based on the stability of their trajectories in the oscillating [electric fields](#) that are applied to the rods.



Principle of operation

- A quadrupole mass filter consists of four parallel metal rods with different charges
- Two opposite rods have an applied + potential and the other two rods have a - potential
- The applied voltages affect the trajectory of ions traveling down the flight path
- For given dc and ac voltages, only ions of a certain mass-to-charge ratio pass through the quadrupole filter and all other ions are thrown out of their original path

Ideally, the rods are [hyperbolic](#). Cylindrical rods with a specific ratio of rod diameter-to-spacing provide an easier-to-manufacture adequate approximation to hyperbolas. Small variations in the ratio have large effects on resolution and peak shape. Different manufacturers choose slightly different ratios to fine-tune operating characteristics in context of anticipated application requirements. In recent decades some manufacturers have produced quadrupole mass spectrometers with true hyperbolic rods.

Applications

These mass spectrometers excel at applications where particular ions of interest are being studied because they can stay tuned on a single ion for extended periods of time. One place where this is useful is in liquid chromatography-mass spectrometry or gas chromatography-mass spectrometry where they serve as exceptionally high specificity detectors. Quadrupole instruments are often reasonably priced and make good multi-purpose instruments.

Detectors

The final element of the mass spectrometer is the detector. The detector records either the charge induced or the current produced when an ion passes by or hits a surface. In a scanning instrument, the signal produced in the detector during the course of the scan versus where the instrument is in the scan (at what m/Q) will produce a mass spectrum, a record of ions as a function of m/Q .

Typically, some type of electron multiplier is used, though other detectors including Faraday cups and ion-to-photon detectors are also used. Because the number of ions leaving the mass analyzer at a particular instant is typically quite small, considerable amplification is often necessary to get a signal. Microchannel plate detectors are commonly used in modern commercial instruments. In FTMS and Orbitraps, the detector consists of a pair of metal surfaces within the mass analyzer/ion trap region which the ions only pass near as they oscillate. No direct current is produced, only a weak AC image current is produced in a circuit between the electrodes. Other inductive detectors have also been used

Separation techniques combined with mass spectrometry

An important enhancement to the mass resolving and mass determining capabilities of mass spectrometry is using it in tandem with chromatographic and other separation techniques.

Gas chromatography

A common combination is gas chromatography-mass spectrometry (GC/MS or GC-MS). In this technique, a gas chromatograph is used to separate different compounds. This stream of separated compounds is fed online into the ion source, a metallic filament to which voltage is applied. This filament emits electrons which ionize the compounds. The ions can then further fragment, yielding predictable patterns. Intact ions and fragments pass into the mass spectrometer's analyzer and are eventually detected.

Liquid chromatography

Similar to gas chromatography MS (GC/MS), liquid chromatography-mass spectrometry (LC/MS or LC-MS) separates compounds chromatographically before they are introduced to the ion source and mass spectrometer. It differs from GC/MS in that the mobile phase is liquid, usually a mixture of water and organic solvents, instead of gas. Most commonly, an electrospray ionization source is used in LC/MS. Other popular and commercially available LC/MS ion sources are atmospheric pressure chemical ionization and atmospheric pressure photoionization. There are also some newly developed ionization techniques like laser spray.

Capillary electrophoresis–mass spectrometry

Capillary electrophoresis–mass spectrometry (CE-MS) is a technique that combines the liquid separation process of capillary electrophoresis with mass spectrometry. CE-MS is typically coupled to electrospray ionization.

Ion mobility

Ion mobility spectrometry-mass spectrometry (IMS/MS or IMMS) is a technique where ions are first separated by drift time through some neutral gas under an applied electrical potential gradient before being introduced into a mass spectrometer. Drift time is a measure of the radius relative to the charge of the ion. The duty cycle of IMS (the time over which the experiment takes place) is longer than most mass spectrometric techniques, such that the mass spectrometer can sample along the course of the IMS separation. This produces data about the IMS separation and the mass-to-charge ratio of the ions in a manner similar to LC/MS.

The duty cycle of IMS is short relative to liquid chromatography or gas chromatography separations and can thus be coupled to such techniques, producing triple modalities such as LC/IMS/MS

Polyacrylamide gel electrophoresis (PAGE)

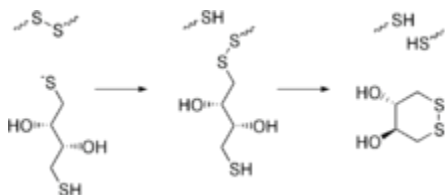
Polyacrylamide gel electrophoresis (PAGE), describes a technique widely used in [biochemistry](#), [forensics](#), [genetics](#), [molecular biology](#) and [biotechnology](#) to separate biological [macromolecules](#), usually [proteins](#) or [nucleic acids](#), according to their [electrophoretic mobility](#). Mobility is a function of the length, conformation and charge of the molecule.

As with all forms of [gel electrophoresis](#), molecules may be run in their [native state](#), preserving the molecules' higher-order structure, or a chemical denaturant may be added to remove this structure and turn the molecule into an unstructured linear chain whose mobility depends only on its length and mass-to-charge ratio. For nucleic acids, [urea](#) is the most commonly used denaturant. For proteins, [sodium dodecyl sulfate](#) (SDS) is an anionic detergent applied to protein samples to linearize proteins and to impart a negative charge to linearized proteins. This procedure is called **SDS-PAGE**. In most proteins, the binding of SDS to the polypeptide chain imparts an even distribution of charge per unit mass, thereby resulting in a fractionation by approximate size during electrophoresis. Proteins that have a greater hydrophobic content, for instance many membrane proteins, and those that interact with [surfactants](#) in their native environment, are intrinsically harder to treat accurately using this method, due to the greater variability in the ratio of bound SDS.

Procedure

Sample preparation

Samples may be any material containing proteins or nucleic acids. These may be biologically derived, for example from prokaryotic or eukaryotic cells, tissues, viruses, environmental samples, or purified proteins. In the case of solid tissues or cells, these are often first broken down mechanically using a blender (for larger sample volumes), using a homogenizer (smaller volumes), by sonicator or by using cycling of high pressure, and a combination of biochemical and mechanical techniques – including various types of filtration and centrifugation – may be used to separate different cell compartments and organelles prior to electrophoresis. Synthetic biomolecules such as oligonucleotides may also be used as analytes.



Reduction of a typical disulfide bond by DTT via two sequential thiol-disulfide exchange reactions.

The sample to analyze is optionally mixed with a chemical denaturant if so desired, usually SDS for proteins or urea for nucleic acids. SDS is an anionic detergent that denatures secondary and non-disulfide-linked tertiary structures, and additionally applies a negative charge to each protein in proportion to its mass. Urea breaks the hydrogen bonds between the base pairs of the nucleic acid, causing the constituent strands to separate. Heating the samples to at least 60 °C further promotes denaturation.[2][3][4][5]

In addition to SDS, proteins may optionally be briefly heated to near boiling in the presence of a reducing agent, such as dithiothreitol (DTT) or 2-mercaptoethanol (beta-mercaptoethanol/BME), which further denatures the proteins by reducing disulfide linkages, thus overcoming some forms of tertiary protein folding, and breaking up quaternary protein structure (oligomeric subunits). This is known as reducing SDS-PAGE.

A tracking dye may be added to the solution. This typically has a higher electrophoretic mobility than the analytes to allow the experimenter to track the progress of the solution through the gel during the electrophoretic run.

Preparing acrylamide gels

The gels typically consist of acrylamide, bisacrylamide, the optional denaturant (SDS or urea), and a buffer with an adjusted pH. The solution may be degassed under a vacuum to prevent the formation of air bubbles during polymerization. Alternatively, butanol may be added to the resolving gel (for proteins) after it is poured, as butanol removes bubbles and makes the surface smooth. A source of free radicals and a stabilizer, such as ammonium persulfate and TEMED are added to initiate polymerization. The polymerization reaction creates a gel because of the added bisacrylamide, which can form cross-links between two acrylamide molecules. The ratio of bisacrylamide to acrylamide can be varied for special purposes, but is generally about 1 part in 35. The acrylamide concentration of the gel can also be varied, generally in the range from 5% to 25%. Lower percentage gels are better for resolving very high molecular weight molecules, while much higher percentages are needed to resolve smaller proteins. Gels are usually polymerized between two glass plates in a gel caster, with a comb

inserted at the top to create the sample wells. After the gel is polymerized the comb can be removed and the gel is ready for electrophoresis.

Electrophoresis

Various buffer systems are used in PAGE depending on the nature of the sample and the experimental objective. The buffers used at the anode and cathode may be the same or different. An electric field is applied across the gel, causing the negatively charged proteins or nucleic acids to migrate across the gel away from the negative electrode (which is the cathode being that this is an electrolytic rather than galvanic cell) and towards the positive electrode (the anode). Depending on their size, each biomolecule moves differently through the gel matrix: small molecules more easily fit through the pores in the gel, while larger ones have more difficulty. The gel is run usually for a few hours, though this depends on the voltage applied across the gel; migration occurs more quickly at higher voltages, but these results are typically less accurate than at those at lower voltages. After the set amount of time, the biomolecules have migrated different distances based on their size. Smaller biomolecules travel farther down the gel, while larger ones remain closer to the point of origin. Biomolecules may therefore be separated roughly according to size, which depends mainly on molecular weight under denaturing conditions, but also depends on higher-order conformation under native conditions. However, certain glycoproteins behave anomalously on SDS gels.

Following electrophoresis, the gel may be stained (for proteins, most commonly with Coomassie Brilliant Blue R-250; for nucleic acids, ethidium bromide; or for either, silver stain), allowing visualization of the separated proteins, or processed further (e.g. Western blot). After staining, different species biomolecules appear as distinct bands within the gel. It is common to run molecular weight size markers of known molecular weight in a separate lane in the gel to calibrate the gel and determine the approximate molecular mass of unknown biomolecules by comparing the distance traveled relative to the marker.

For proteins, SDS-PAGE is usually the first choice as an assay of purity due to its reliability and ease. The presence of SDS and the denaturing step make proteins separate, approximately based on size, but aberrant migration of some proteins may occur. Different proteins may also stain differently, which interferes with quantification by staining. PAGE may also be used as a preparative technique for the purification of proteins. For

example, quantitative preparative native continuous polyacrylamide gel electrophoresis (QPNC-PAGE) is a method for separating native metalloproteins in complex biological matrices.

Chemical ingredients and their roles

Polyacrylamide gel (PAG) had been known as a potential embedding medium for sectioning tissues as early as 1964, and two independent groups employed PAG in electrophoresis in 1959.[11][12] It possesses several electrophoretically desirable features that make it a versatile medium. It is a synthetic, thermo-stable, transparent, strong, chemically relatively inert gel, and can be prepared with a wide range of average pore sizes.[13] The pore size of a gel is determined by two factors, the total amount of acrylamide present (%T) (T = Total concentration of acrylamide and bisacrylamide monomer) and the amount of cross-linker (%C) (C = bisacrylamide concentration). Pore size decreases with increasing %T; with cross-linking, 5%C gives the smallest pore size. Any increase or decrease in %C from 5% increases the pore size, as pore size with respect to %C is a parabolic function with vertex as 5%C. This appears to be because of non-homogeneous bundling of polymer strands within the gel. This gel material can also withstand highvoltage gradients, is amenable to various staining and destaining procedures, and can be digested to extract separated fractions or dried for autoradiography and permanent recording.

Components

Chemical buffer Stabilizes the pH value to the desired value within the gel itself and in the electrophoresis buffer. The choice of buffer also affects the electrophoretic mobility of the buffer counterions and thereby the resolution of the gel. The buffer should also be unreactive and not modify or react with most proteins. Different buffers may be used as cathode and anode buffers, respectively, depending on the application. Multiple pH values may be used within a single gel, for example in DISC electrophoresis. Common buffers in PAGE include Tris, Bis-Tris, or imidazole.

Counter ion balance the intrinsic charge of the buffer ion and also affect the electric field strength during electrophoresis. Highly charged and mobile ions are often avoided in SDS-PAGE cathode buffers, but may be included in the gel itself, where it migrates ahead of the protein. In applications such as DISC SDS-PAGE the pH values within the gel may vary to change the average charge of the counterions during the run to improve resolution. Popular counterions are glycine and tricine. Glycine has been used as the source of trailing ion or slow ion because its

pKa is 9.69 and mobility of glycinate are such that the effective mobility can be set at a value below that of the slowest known proteins of net negative charge in the pH range. The minimum pH of this range is approximately 8.0.

Acrylamide (C_3H_5NO ; mW: 71.08). When dissolved in water, slow, spontaneous autopolymerization of acrylamide takes place, joining molecules together by head on tail fashion to form long single-chain polymers. The presence of a free radical-generating system greatly accelerates polymerization. This kind of reaction is known as Vinyladdition polymerisation. A solution of these polymer chains becomes viscous but does not form a gel, because the chains simply slide over one another. Gel formation requires linking various chains together. Acrylamide is a neurotoxin. It is also essential to store acrylamide in a cool dark and dry place to reduce autopolymerisation and hydrolysis.

Bisacrylamide (**N,N'-Methylenebisacrylamide**) ($C_7H_{10}N_2O_2$; mW: 154.17). Bisacrylamide is the most frequently used cross linking agent for polyacrylamide gels. Chemically it can be thought of as two acrylamide molecules coupled head to head at their non-reactive ends. Bisacrylamide can crosslink two polyacrylamide chains to one another, thereby resulting in a gel.

Sodium Dodecyl Sulfate (SDS) ($C_{12}H_{25}NaO_4S$; mW: 288.38). (only used in denaturing protein gels) SDS is a strong detergent agent used to denature native proteins to unfolded, individual polypeptides. When a protein mixture is heated to 100 °C in presence of SDS, the detergent wraps around the polypeptide backbone. It binds to polypeptides in a constant weight ratio of 1.4 g SDS/g of polypeptide. In this process, the intrinsic charges of polypeptides becomes negligible when compared to the negative charges contributed by SDS. Thus polypeptides after treatment become rod-like structures possessing a uniform charge density, that is same net negative charge per unit weight. The electrophoretic mobilities of these proteins is a linear function of the logarithms of their molecular weights.

Without SDS, different proteins with similar molecular weights would migrate differently due to differences in mass-charge ratio, as each protein has an isoelectric point and molecular weight particular to its primary structure. This is known as Native PAGE. Adding SDS solves this problem, as it binds to and unfolds the protein, giving a near uniform negative charge along the length of the polypeptide.

Urea ($CO(NH_2)_2$; mW: 60.06). Urea is a chaotropic agent that increases the entropy of the system by interfering with intramolecular interactions mediated by non-covalent forces such

as hydrogen bonds and van der Waals forces. Macromolecular structure is dependent on the net effect of these forces, therefore it follows that an increase in chaotropic solutes denatures macromolecules,

Ammonium persulfate (APS) ($\text{N}_2\text{H}_8\text{S}_2\text{O}_8$; mW: 228.2). APS is a source of free radicals and is often used as an initiator for gel formation. An alternative source of free radicals is riboflavin, which generated free radicals in a photochemical reaction.

TEMED (N, N, N', N'-tetramethylethylenediamine) ($\text{C}_6\text{H}_{16}\text{N}_2$; mW: 116.21). TEMED stabilizes free radicals and improves polymerization. The rate of polymerisation and the properties of the resulting gel depend on the concentrations of free radicals. Increasing the amount of free radicals results in a decrease in the average polymer chain length, an increase in gel turbidity and a decrease in gel elasticity. Decreasing the amount shows the reverse effect. The lowest catalytic concentrations that allow polymerisation in a reasonable period of time should be used. APS and TEMED are typically used at approximately equimolar concentrations in the range of 1 to 10 mM.

Chemicals for processing and visualization

The following chemicals and procedures are used for processing of the gel and the protein samples visualized in it:

Tracking dye. As proteins and nucleic acids are mostly colorless, their progress through the gel during electrophoresis cannot be easily followed. Anionic dyes of a known electrophoretic mobility are therefore usually included in the PAGE sample buffer. A very common tracking dye is Bromophenol blue (BPB, 3',3'',5',5'' tetra bromophenol sulfonphthalein). This dye is coloured at alkali and neutral pH and is a small negatively charged molecule that moves towards the anode. Being a highly mobile molecule it moves ahead of most proteins. As it reaches the anodic end of the electrophoresis medium electrophoresis is stopped. It can weakly bind to some proteins and impart a blue colour. Other common tracking dyes are xylene cyanol, which has lower mobility, and Orange G, which has a higher mobility.

Loading aids. Most PAGE systems are loaded from the top into wells within the gel. To ensure that the sample sinks to the bottom of the gel, sample buffer is supplemented with additives that increase the density of the sample. These additives should be non-ionic and non-reactive towards proteins to avoid interfering with electrophoresis. Common additives are glycerol and sucrose.

Coomassie Brilliant Blue R-250 (CBB)($C_{45}H_{44}N_3NaO_7S_2$; mW: 825.97). CBB is the most popular protein stain. It is an anionic dye, which non-specifically binds to proteins. The structure of CBB is predominantly non-polar, and it is usually used in methanolic solution acidified with acetic acid. Proteins in the gel are fixed by acetic acid and simultaneously stained. The excess dye incorporated into the gel can be removed by destaining with the same solution without the dye. The proteins are detected as blue bands on a clear background. As SDS is also anionic, it may interfere with staining process. Therefore, large volume of staining solution is recommended, at least ten times the volume of the gel.

Ethidium bromide (EtBr) is the traditionally most popular nucleic acid stain.

Silver staining. Silver staining is used when more sensitive method for detection is needed, as classical Coomassie Brilliant Blue staining can usually detect a 50 ng protein band, Silver staining increases the sensitivity typically 50 times. The exact chemical mechanism by which this happens is still largely unknown. Silver staining was introduced by Kerenyi and Gallyas as a sensitive procedure to detect trace amounts of proteins in gels. The technique has been extended to the study of other biological macromolecules that have been separated in a variety of supports. Many variables can influence the colour intensity and every protein has its own staining characteristics; clean glassware, pure reagents and water of highest purity are the key points to successful staining. Silver staining was developed in the 14th century for colouring the surface of glass. It has been used extensively for this purpose since the 16th century. The colour produced by the early silver stains ranged between light yellow and an orange-red. Camillo Golgi perfected the silver staining for the study of the nervous system. Golgi's method stains a limited number of cells at random in their entirety.

Western Blotting is a process by which proteins separated in the acrylamide gel are electrophoretically transferred to a stable, manipulable membrane such as nitrocellulose, nylon, or PVDF membrane. It is then possible to apply immunochemical techniques to visualise the transferred proteins, as well as accurately identify relative increases or decreases of the protein of interest. For more, see Western Blot

PROTEIN MICROARRAY

A **protein microarray** (or **protein chip**) is a high-throughput method used to track the interactions and activities of proteins, and to determine their function, and determining function on a large scale. Its main advantage lies in the fact that large numbers of proteins can be tracked in parallel. The chip consists of a support surface such as a glass slide, nitrocellulose membrane, bead, or microtitre plate, to which an array of capture proteins is bound.^[2] Probe molecules, typically labeled with a fluorescent dye, are added to the array. Any reaction between the probe and the immobilised protein emits a fluorescent signal that is read by a laser scanner.^[3] Protein microarrays are rapid, automated, economical, and highly sensitive, consuming small quantities of samples and reagents. The concept and methodology of protein microarrays was first introduced and illustrated in antibody microarrays (also referred to as antibody matrix) in 1983 in a scientific publication^[5] and a series of patents.^[6] The high-throughput technology behind the protein microarray was relatively easy to develop since it is based on the technology developed for DNA microarrays, which have become the most widely used microarrays.

Motivation for development

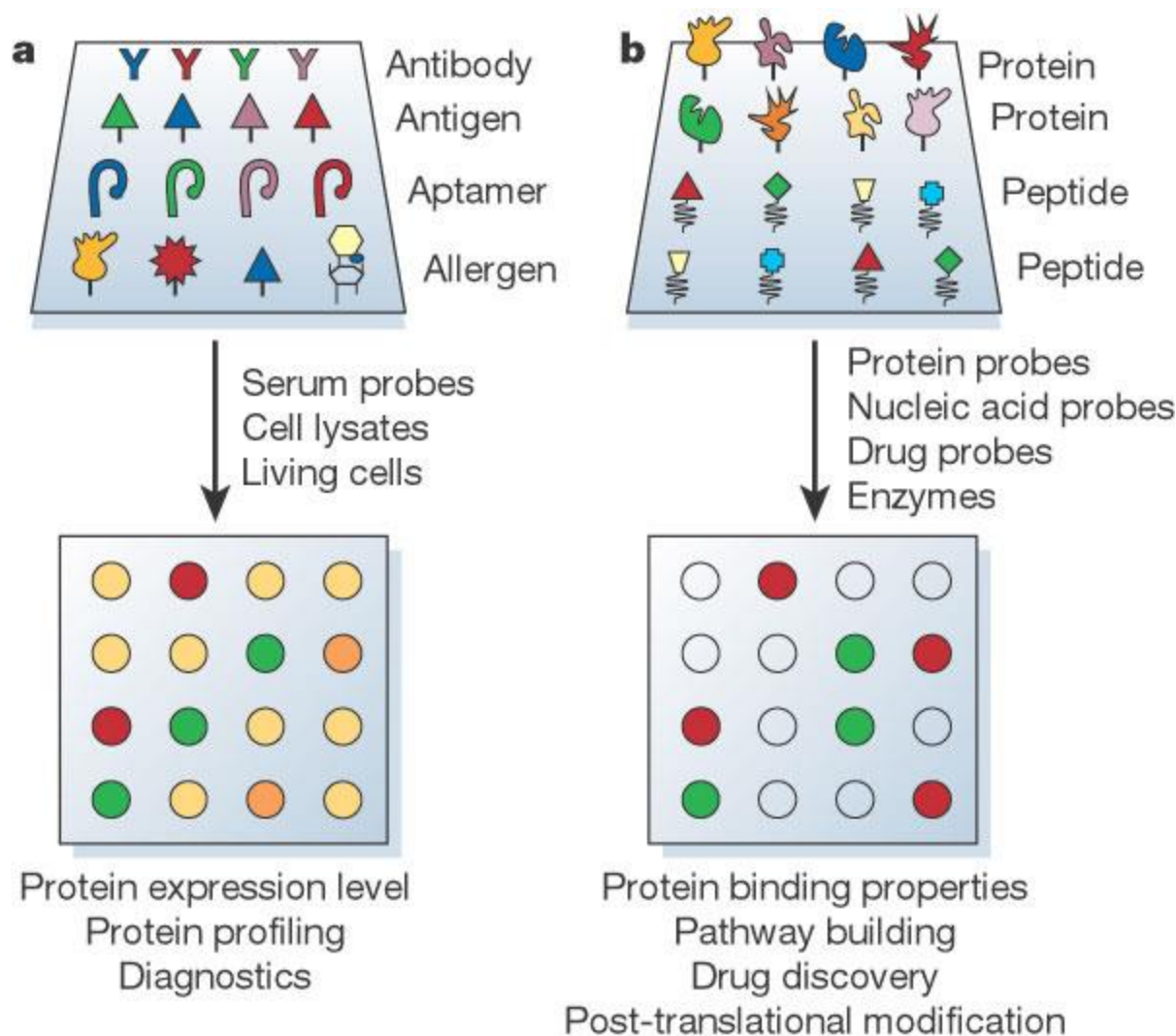
Protein microarrays were developed due to the limitations of using DNA microarrays for determining gene expression levels in proteomics. The quantity of mRNA in the cell often doesn't reflect the expression levels of the proteins they correspond to. Since it is usually the protein, rather than the mRNA, that has the functional role in cell response, a novel approach was needed. Additionally post-translational modifications, which are often critical for determining protein function, are not visible on DNA microarrays.^[8] Protein microarrays replace traditional proteomics techniques such as 2D gel electrophoresis or chromatography, which were time consuming, labor-intensive and ill-suited for the analysis of low abundant proteins.

Making the array

The proteins are arrayed onto a solid surface such as microscope slides, membranes, beads or microtitre plates. The function of this surface is to provide a support onto which proteins can be immobilized. It should demonstrate maximal binding properties, whilst maintaining the protein in its native conformation so that its binding ability is retained. Microscope slides made of glass or silicon are a popular choice since they are compatible with the easily obtained robotic arrayers and laser scanners that have been developed for DNA

microarray technology. Nitrocellulose film slides are broadly accepted as the highest protein binding substrate for protein microarray applications.

The chosen solid surface is then covered with a coating that must serve the simultaneous functions of immobilising the protein, preventing its denaturation, orienting it in the appropriate direction so that its binding sites are accessible, and providing a hydrophilic environment in which the binding reaction can occur. In addition, it also needs to display minimal non-specific binding in order to minimize background noise in the detection systems. Furthermore, it needs to be compatible with different detection systems. Immobilising agents include layers of aluminium or gold, hydrophilic polymers, and polyacrylamide gels, or treatment with amines, aldehyde or epoxy. Thin-film technologies like physical vapour deposition (PVD) and chemical vapour deposition (CVD) are employed to apply the coating to the support surface.



Types of Protein Microarray and Antibody Chips

Protein Chips and Antibody Microarrays - Glass Slides, Microwell/Nanowell

Two major protein chip formats are now used, including glass slides and nanowells. Due to the large amounts of slide adherence methods only the most common and major types will be mentioned here.

It is important that protein chips retain proteins in an active state at high densities, are compatible with most commercial arrayers and scanners, and can be printed in such a fashion that the proteins remain in a moisturized environment.

Glass Slide Chips

Glass slides have the advantage that they are compatible with standard microarray equipment and detection equipment used for DNA chips. They are also inexpensive. The majority of studies are now using glass slides. However, they have a high evaporation rate and are susceptible to possible cross-contamination.

The first strategies for the creation of protein arrays on glass were developed by Mirzabekov et al.. Arrays were produced by immobilizing proteins in tiny gel pockets that were attached to the glass surface. A variety of immunoassays, antigen detection, and enzymatic assays were carried out. Due to the 3D matrix structure, protein immobilization was very efficient.

In another study, proteins were attached to a glass surface activated with a crosslinking agent that reacts with primary amines. Proteins were spotted in a 40% glycerol solution which keeps proteins in a wet environment and prevents dehydration. To determine that their protein microarrays were feasible for biochemical assays, they tested three known protein-protein interactions, three known kinase-substrate reactions, and three known protein-ligand interactions using fluorophore-tagged proteins, radiolabeled ATP, and synthetic ligands coupled to fluorescently labeled bovine serum albumin (BSA), respectively. In most of the experiments, small numbers of proteins were spotted however in one experiment they identified a single spot within an array of 10,000 spots containing one other protein. This work demonstrated the potential of protein microarrays in large-scale biochemical assays however their study analyzed a very small number of proteins and novel activities were not identified.

Another study used glass slides for the detection of several protein antigens. Analytes were spotted onto a treated glass surface at high density using a hand-spotting device or a microarray robot. The spotted analytes were detected using antibodies attached to an

oligonucleotide primer and a rolling circle amplification reaction. The technique had a high sensitivity, a wide dynamic range, and excellent spot-to-spot reproducibility. Most groups now directly array proteins and antibodies onto plain glass slides and spotting is carried out in a humidity-controlled environment.

Microwell/Nanowell Chips

Compatible with standard microarray and detection equipment however alignment is required. This method is versatile for solution-based assays and multiple-component reactions. Evaporation is reduced and there is no cross-contamination. Also these chips are relatively inexpensive.

Zhu et al., fabricated an open structure, namely nanowells on a polydimethylsiloxane (PDMS) surface supported by the standard glass slides (36). These chips consist of an array of microwells in a disposable silicone elastomer, poly(dimethylsiloxane) (PDMS). Microwell arrays allow small volumes of different analytes to be densely packed on a single chip, yet remain physically segregated during subsequent batch processing. Proteins were covalently attached to the wells using a crosslinker 3-glycidoxypentyltrimethoxysilane (GPTS).

Captured molecules can be easily recovered from the nanowells. When covered with gold in the nanowells, it is expected that high-throughput mass spectrometry and surface plasmon resonance analyses can be performed. The greatest disadvantage of this technique is that specialized equipment is required to load the nanowells at high density.

An aqueous environment is essential at all stages of array manufacture and operation to prevent protein denaturation. Therefore, sample buffers contain a high percent of glycerol (to lower the freezing point), and the humidity of the manufacturing environment is carefully regulated. Microwells have the dual advantage of providing an aqueous environment while preventing cross-contamination between samples.

Attachment

To attach proteins to a solid surface, the surface of the substrate has to be modified to achieve maximum binding capacity. The proteins are attached to the chip on a protein attachment layer. This layer is typically an organic film which varies with the nature of the application. A variety of materials have been studied including agarose (39), dextran-based hydrogel, porous polyacrylamide hydrogel hydrophilic polymers and polyamino acids.

A convenient attachment method used nitrocellulose-membrane or poly-L-Lysine coated glass such that proteins could be passively absorbed onto the surface through non-specific interactions. The attached proteins bind onto the surface in random orientations and can be washed off under stringent washing conditions. However, the noise level is usually higher because of the non-specific absorption/adsorption.

A more specific and stronger attachment is achieved by creating reactive surfaces on glass that can covalently cross-link to proteins. A bifunctional silane cross-linker is used to form a self-assembled monolayer (SAM), which has one functional group that reacts with the hydroxyl groups on glass the glass surface, and another group which is free to react with primary amine groups of proteins or can be further chemically modified to reach maximum specificity. Another variation is gold-coated glass. The advantage of gold-coated chips is that SPR and mass spectrometry can be integrated as detection methods to monitor the dynamics of the reaction, and to identify the captured molecules.

The above-mentioned covalent cross-linking approaches however have a disadvantage. Due to the fact that reactive ligands also exist in the side chains of proteins it is possible that their random attachment may alter the native confirmation of proteins, reduce the activity of proteins, or make them inaccessible to probes (8,27).

In order to orient proteins uniformly away from the surface of the chip, proteins may be fused with a high-affinity tag at their amino or carboxy termini. With this method, immobilized proteins/antibodies are more likely to remain in their native conformation, thus allowing the analytes efficient access to the active sites of the proteins. This method was first successfully demonstrated with the attachment of 5800 fusion proteins containing a His tag onto a nickel-coated glass slide (26). Other affinity methods such as glutathione/GST have also been used.

Streptavidin based immobilization methods have been also widely employed to attach any biotinylated biological element to the array surface.

The chip support material is important because proteins are highly sensitive to physiochemical properties. For example, polar arrays are chemically treated to bind to hydrophilic proteins however such surfaces are unsuitable for cell membrane proteins (e.g. G-protein coupled receptors) as they possess hydrophobic domains

Proteins do not behave like nucleic acids, and different proteins will behave in different ways when exposed to the same surface chemistry. Different types a surface

chemistries will thus promote the retention of some proteins and cause denaturation or loss of activity of others. Therefore, the proper choice of surface chemistry is important as this will allow immobilized proteins of diverse types to retain their secondary and tertiary structures, and thus their biological activity. This problem is magnified when the number of different spots on the chip increases, as there may be 100 different ways to immobilize 100 different proteins in order to obtain proper folding and function of all the proteins. Also a significant problem is because the functions of most proteins are currently unknown, so there is no method to actually test whether they are still functional on the chip

In the most common type of protein array, robots place large numbers of proteins or their ligands onto a coated solid support in a pre-defined pattern. This is known as robotic contact printing or robotic spotting. Another fabrication method is ink-jetting, a drop-on-demand, non-contact method of dispersing the protein polymers onto the solid surface in the desired pattern.^[9] Piezoelectric spotting is a similar method to ink-jet printing. The printhead moves across the array, and at each spot uses electric stimulation to deliver the protein molecules onto the surface via tiny jets. This is also a non-contact process.^[10] Photolithography is a fourth method of arraying the proteins onto the surface. Light is used in association with photomasks, opaque plates with holes or transparencies that allow light to shine through in a defined pattern. A series of chemical treatments then enables deposition of the protein in the desired pattern upon the material underneath the photomask.

The capture molecules arrayed on the solid surface may be antibodies, antigens, aptamers (nucleic acid-based ligands), affibodies (small molecules engineered to mimic monoclonal antibodies), or full length proteins. Sources of such proteins include cell-based expression systems for recombinant proteins, purification from natural sources, production in vitro by cell-free translation systems, and synthetic methods for peptides. Many of these methods can be automated for high throughput production but care must be taken to avoid conditions of synthesis or extraction that result in a denatured protein which, since it no longer recognizes its binding partner, renders the array useless.

Proteins are highly sensitive to changes in their microenvironment. This presents a challenge in maintaining protein arrays in a stable condition over extended periods of time. In situ methods involve on-chip synthesis of proteins as and when required, directly from the DNA using cell-free protein expression systems. Since DNA is a highly stable molecule it does not

deteriorate over time and is therefore suited to long-term storage. This approach is also advantageous in that it circumvents the laborious and often costly processes of separate protein purification and DNA cloning, since proteins are made and immobilised simultaneously in a single step on the chip surface. Examples of In situ techniques are PISA (protein in situ array), NAPPA (nucleic acid programmable protein array) and DAPA (DNA array to protein array).

Types of protein arrays

There are three types of protein microarrays that are currently used to study the biochemical activities of proteins.

Analytical microarrays are also known as capture arrays. In this technique, a library of antibodies, aptamers or affibodies is arrayed on the support surface. These are used as capture molecules since each binds specifically to a particular protein. The array is probed with a complex protein solution such as a cell lysate. Analysis of the resulting binding reactions using various detection systems can provide information about expression levels of particular proteins in the sample as well as measurements of binding affinities and specificities. This type of microarray is especially useful in comparing protein expression in different solutions. For instance the response of the cells to a particular factor can be identified by comparing the lysates of cells treated with specific substances or grown under certain conditions with the lysates of control cells. Another application is in the identification and profiling of diseased tissues.

Functional protein microarrays (also known as target protein arrays) are constructed by immobilising large numbers of purified proteins and are used to identify protein-protein, protein-DNA, protein-RNA, protein-phospholipid, and protein-small molecule interactions, to assay enzymatic activity and to detect antibodies and demonstrate their specificity. They differ from analytical arrays in that functional protein arrays are composed of arrays containing full-length functional proteins or protein domains. These protein chips are used to study the biochemical activities of the entire proteome in a single experiment.

Reverse phase protein microarray (RPPA) involve complex samples, such as tissue lysates. Cells are isolated from various tissues of interest and are lysed. The lysate is arrayed onto the microarray and probed with antibodies against the target protein of interest. These antibodies are typically detected with chemiluminescent, fluorescent or colorimetric assays. Reference peptides are printed on the slides to allow for protein quantification of the sample lysates. RPAs allow for

the determination of the presence of altered proteins or other agents that may be the result of disease. Specifically, post-translational modifications, which are typically altered as a result of disease can be detected using RPAs.

Detection

Protein array detection methods must give a high signal and a low background. The most common and widely used method for detection is fluorescence labeling which is highly sensitive, safe and compatible with readily available microarray laser scanners. Other labels can be used, such as affinity, photochemical or radioisotope tags. These labels are attached to the probe itself and can interfere with the probe-target protein reaction.

Therefore, a number of label free detection methods are available, such as surface plasmon resonance (SPR), carbon nanotubes, carbon nanowire sensors (where detection occurs via changes in conductance) and microelectromechanical system (MEMS) cantilevers. All these label free detection methods are relatively new and are not yet suitable for high-throughput protein interaction detection; however, they do offer much promise for the future.

There are a few label free methods of detecting the bound molecules. One method is to use MALDI MS to ionize the molecules attached and determine their nature by their mass spectrum. This is very useful with protein-protein interactions when the other possible interaction proteins is unknown. A second method is to use surface plasmon resonance. This method measures difference in the index of refraction which will change when a large molecule binds to another. In this case the protein array is placed on a thin gold film. A polarized collimated white light is shown on the gold film and the reflected light is measured through a polarized filter. If something is bound there will be a change in the light. This method though is a 100 times less sensitive than fluorescence. A system using carbon nanotubes and nanowires looks for changes in the conductance through these structures. Lastly are micromechanical systems which are now being developed.

Protein quantitation on nitrocellulose coated glass slides can use near-IR fluorescent detection. This limits interferences due to auto-fluorescence of the nitrocellulose at the UV wavelengths used for standard fluorescent detection probes.^[13]

Applications

There are five major areas where protein arrays are being applied: diagnostics, proteomics, protein functional analysis, antibody characterization, and treatment development

- Diagnostics involves the detection of antigens and antibodies in blood samples; the profiling of sera to discover new disease biomarkers; the monitoring of disease states and responses to therapy in personalized medicine; the monitoring of environment and food.
- Proteomics pertains to protein expression profiling i.e. which proteins are expressed in the lysate of a particular cell.
- Protein functional analysis is the identification of protein-protein interactions (e.g. identification of members of a protein complex), protein-phospholipid interactions, small molecule targets, enzymatic substrates (particularly the substrates of kinases) and receptor ligands.
- Antibody characterization is characterizing cross-reactivity, specificity and mapping epitopes.
- Treatment development involves the development of antigen-specific therapies for autoimmunity, cancer and allergies; the identification of small molecule targets that could potentially be used as new drugs.

Challenges

Despite the considerable investments made by several companies, proteins chips have yet to flood the market. Manufacturers have found that proteins are actually quite difficult to handle. A protein chip requires a lot more steps in its creation than does a DNA chip.

- 1) finding a surface and a method of attachment that allows the proteins to maintain their secondary or tertiary structure and thus their biological activity and their interactions with other molecules,
- 2) producing an array with a long shelf life so that the proteins on the chip do not denature over a short time,
- 3) identifying and isolating antibodies or other capture molecules against every protein in the human genome,
- 4) quantifying the levels of bound protein while assuring sensitivity and avoiding background noise,
- 5) extracting the detected protein from the chip in order to further analyze it,
- 6) reducing non-specific binding by the capture agents,
- 7) the capacity of the chip must be sufficient to allow as complete a representation of the proteome to be visualized as possible; abundant proteins overwhelm the detection of less

Protein purification and separation

abundant proteins such as signaling molecules and receptors, which are generally of more therapeutic interest

Post Translational Modifications

Post-translational modification (PTM) refers to the covalent and generally enzymatic modification of proteins during or after protein biosynthesis. Proteins are synthesized by ribosomes translating mRNA into polypeptide chains, which may then undergo PTM to form the mature protein product. PTMs are important components in cell signaling. Most of the proteins that are translated from mRNA undergo chemical modifications before becoming functional in different body cells. The modifications collectively, are known as post-translational modifications. The protein post translational modifications play a crucial role in generating the heterogeneity in proteins and also help in utilizing identical proteins for different cellular functions in different cell types. How a particular protein sequence will act in most of the eukaryotic organisms is regulated by these post translational modifications.

Post-translational modifications can occur on the amino acid side chains or at the protein's C- or N- termini. They can extend the chemical repertoire of the 20 standard amino acids by introducing new functional groups such as phosphate, acetate, amide groups, or methyl groups. Phosphorylation is a very common mechanism for regulating the activity of enzymes and is the most common post-translational modification.[2] Many eukaryotic proteins also have carbohydrate molecules attached to them in a process called glycosylation, which can promote protein folding and improve stability as well as serving regulatory functions. Attachment of lipid molecules, known as lipidation, often targets a protein or part of a protein to the cell membrane.

Other forms of post-translational modification consist of cleaving peptide bonds, as in processing a propeptide to a mature form or removing the initiator methionine residue. The formation of disulfide bonds from cysteine residues may also be referred to as a post-translational modification. For instance, the peptide hormone insulin is cut twice after disulfide bonds are formed, and a propeptide is removed from the middle of the chain; the resulting protein consists of two polypeptide chains connected by disulfide bonds.

Some types of post-translational modification are consequences of oxidative stress. Carbonylation is one example that targets the modified protein for degradation and can result in the formation of protein aggregates. Specific amino acid modifications can be used as biomarkers indicating oxidative damage.

Sites that often undergo post-translational modification are those that have a functional group that can serve as a nucleophile in the reaction: the hydroxyl groups of serine, threonine, and tyrosine; the amine forms of lysine, arginine, and histidine; the thiolate anion of cysteine; the carboxylates of aspartate and glutamate; and the N- and C-termini. In addition, although the amides of asparagine and glutamine are weak nucleophiles, both can serve as attachment points for glycans. Rarer modifications can occur at oxidized methionines and at some methylenes in side chains.

Post-translational modification of proteins can be experimentally detected by a variety of techniques, including mass spectrometry, Eastern blotting, and Western blotting.

Post translational modifications occurring at the peptide terminus of the amino acid chain play an important role in translocating them across biological membranes. These include secretory proteins in prokaryotes and eukaryotes and also proteins that are intended to be incorporated in various cellular and organelle membranes such as lysosomes, chloroplast, mitochondria and plasma membranes.

Expression of proteins is important in diseased conditions. Post translational modifications play an important part in modifying the end product of expression and contribute towards biological processes and diseased conditions. The amino terminal sequences are removed by proteolytic cleavage when the proteins cross the membranes. These amino terminal sequences target the proteins for transporting them to their actual point of action in the cell.

Protein post translational modifications may happen in several ways. Some of them are listed below:

Glycosylation: Many proteins, particularly in eukaryotic cells, are modified by the addition of carbohydrates, a process called glycosylation. Glycosylation in proteins results in addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine. Software for studying glycosylation by glycan structure prediction.

Protein glycosylation is acknowledged as one of the major post-translational modifications, with significant effects on protein folding, conformation, distribution, stability and activity. Glycosylation encompasses a diverse selection of sugar-moiety additions to proteins that ranges from simple monosaccharide modifications of nuclear transcription factors to highly complex branched polysaccharide changes of cell surface receptors. Carbohydrates in the form

of asparagine-linked (N-linked) or serine/threonine-linked (O-linked) oligosaccharides are major structural components of many cell surface and secreted proteins.

Acetylation: the addition of an acetyl group, usually at the N-terminus of the protein. N-acetylation, or the transfer of an acetyl group to nitrogen, occurs in almost all eukaryotic proteins through both irreversible and reversible mechanisms. N-terminal acetylation requires the cleavage of the N-terminal methionine by methionine aminopeptidase (MAP) before replacing the amino acid with an acetyl group from acetyl-CoA by N-acetyltransferase (NAT) enzymes. This type of acetylation is co-translational, in that N-terminus is acetylated on growing polypeptide chains that are still attached to the ribosome. While 80-90% of eukaryotic proteins are acetylated in this manner, the exact biological significance is still unclear

Acetylation at the ϵ -NH₂ of lysine (termed lysine acetylation) on histone N-termini is a common method of regulating gene transcription. Histone acetylation is a reversible event that reduces chromosomal condensation to promote transcription, and the acetylation of these lysine residues is regulated by transcription factors that contain histone acetyltransferase (HAT) activity. While transcription factors with HAT activity act as transcription co-activators, histone deacetylase (HDAC) enzymes are co-repressors that reverse the effects of acetylation by reducing the level of lysine acetylation and increasing chromosomal condensation.

Sirtuins (silent information regulator) are a group of NAD-dependent deacetylases that target histones. As their name implies, they maintain gene silencing by hypoacetylating histones and have been reported to aid in maintaining genomic stability

While acetylation was first detected in histones, cytoplasmic proteins have been reported to also be acetylated, and therefore acetylation seems to play a greater role in cell biology than simply transcriptional regulation (9). Furthermore, crosstalk between acetylation and other post-translational modifications, including phosphorylation, ubiquitination and methylation, can modify the biological function of the acetylated protein.

Protein acetylation can be detected by chromosome immunoprecipitation (ChIP) using acetyllysine-specific antibodies or by mass spectrometry, where an increase in histone by 42 mass units represents a single acetylation.

Alkylation: The addition of an alkyl group (e.g. methyl, ethyl).

S-Nitrosylation Nitric oxide (NO) is produced by three isoforms of nitric oxide synthase (NOS) and is a chemical messenger that reacts with free cysteine residues to form S-nitrothiols (SNOs). S-nitrosylation is a critical PTM used by cells to stabilize proteins, regulate gene expression and provide NO donors, and the generation, localization, activation and catabolism of SNOs are tightly regulated.

S-nitrosylation is a reversible reaction, and SNOs have a short half life in the cytoplasm because of the host of reducing enzymes, including glutathione (GSH) and thioredoxin, that denitrosylate proteins. Therefore, SNOs are often stored in membranes, vesicles, the interstitial space and lipophilic protein folds to protect them from denitrosylation (5). For example, caspases, which mediate apoptosis, are stored in the mitochondrial intermembrane space as SNOs. In response to extra- or intracellular cues, the caspases are released into the cytoplasm, and the highly reducing environment rapidly denitrosylates the proteins, resulting in caspase activation and the induction of apoptosis.

S-nitrosylation is not a random event, and only specific cysteine residues are S-nitrosylated. Because proteins may contain multiple cysteines and due to the labile nature of SNOs, S-nitrosylated cysteines can be difficult to detect and distinguish from non-S-nitrosylated amino acids. The biotin switch assay, developed by Jaffrey et al., is a common method of detecting SNOs, and the steps of the assay are listed below:

- All free cysteines are blocked.
- All remaining cysteines (presumably only those that are denitrosylated) are denitrosylated.
- The now-free thiol groups are then biotinylated.
- Biotinylated proteins are detected by SDS-PAGE and Western blot analysis or mass spectrometry

Methylation: The addition of a methyl group, usually at lysine or arginine residues. (This is a type of alkylation.) The transfer of one-carbon methyl groups to nitrogen or oxygen (N- and O-methylation, respectively) to amino acid side chains increases the hydrophobicity of the protein and can neutralize a negative amino acid charge when bound to carboxylic acids. Methylation is mediated by methyltransferases, and S-adenosyl methionine (SAM) is the primary methyl group donor. Methylation occurs so often that SAM has been suggested to be the most-used substrate in enzymatic reactions after ATP (4). Additionally, while N-methylation is irreversible, O-

methylation is potentially reversible. Methylation is a well-known mechanism of epigenetic regulation, as histone methylation and demethylation influences the availability of DNA for transcription. Amino acid residues can be conjugated to a single methyl group or multiple methyl groups to increase the effects of modification.

Biotinylation: Acylation of conserved lysine residues with a biotin appendage.

Glutamylolation: Covalent linkage of glutamic acid residues to tubulin and some other proteins.

Glycylation: Covalent linkage of one to more than 40 glycine residues to the tubulin C-terminal tail of the amino acid sequence.

Isoprenylation: The addition of an isoprenoid group (e.g. farnesol and geranylgeraniol).

Lipoylation: The attachment of a lipoate functionality.

Phosphopantetheinylation, The addition of a 4'-phosphopantetheinyl moiety from coenzyme A, as in fatty acid, polyketide, non-ribosomal peptide and leucine biosynthesis.

Phosphorylation, the addition of a phosphate group, usually to serine, tyrosine, threonine or histidine.

Sulfation: The addition of a sulfate group to a tyrosine.

Selenation

C-terminal amidation

Cellular Sites of Major PTM's

Site	Modification
Mitochondria/Chloroplasts	Cleavage of Signal Peptides
Golgi Apparatus	Modification of N-glycosyl groups, O-glycosylation with GalNAc
Secretory Vesicles/Granules	Amidation of C-terminus Proteolytic processing of some precursors
ER	Cleavage of signal peptides Core glycosylation of Asn residues Addition of palmitoyl and glycosyl-phosphatidylinositol Hydroxylation of Pro/Lys in procollagen Disulfide bond formation

Functions Enabled by Posttranslational Modification

- Alterations in local folding of proteins
 - E.g. transitions between unstructured and structured regions
 - Generation of charge pairs
- Marking proteins for degradation
 - Proteasome targeting
- Marking chromatin for transcriptional regulation
- Changing the intracellular or extracellular addresses of proteins
 - Signal peptides direct proteins to ...
 - Plasma membrane
 - Secretory pathway
 - Mitochondria
 - Cytosol
- Inactive apo to active holo forms of enzymes

ANALYSIS OF POSTTRANSLATIONAL MODIFICATIONS OF PROTEINS BY TANDEM MASS SPECTROMETRY

INTRODUCTION

Determination of posttranslational modifications (PTM) of proteins is fundamental in elucidation of the intricate processes that govern cellular events, like cell division, growth, and differentiation. The term PTM denotes changes in the polypeptide chain as a result of either the addition or removal of distinct chemical moieties to amino acid residues, proteolytic processing of the protein termini, or the introduction of covalent crosslinks between domains of the protein. PTMs are involved in most cellular processes including the maintenance of protein structure and integrity, regulation of metabolism and defense processes, and in cellular recognition events and morphology changes. Analysis of PTMs presents a number of challenges to protein and proteomics researchers, and efficient and sensitive methods for detection of PTMs are required. Traditionally, PTMs have been identified by Edman degradation, amino acid analysis, isotopic labeling, or immunochemistry. Within recent years, mass spectrometry (MS) has proven to be extremely useful in PTM discovery. The presence of covalent modifications in proteins affects the molecular weight of the modified amino acids, and the mass increment or deficit can be detected by MS (Table 1). MS has several advantages for characterization of PTMs, including (i) very high sensitivity; (ii) ability to identify the site of PTM; (iii) discovery of novel PTMs; (iv) capability to identify PTMs in complex mixtures of proteins; and finally (v) the ability to quantify the relative changes in PTM occupancy at distinct sites. None of the other techniques provide all these features. In this review we describe the utility of tandem mass spectrometry (MS/MS) for the determination of PTMs and provide selected examples of recent modification-specific proteomic studies.

MS IN PROTEOMICS

MS is now widely used in protein biochemistry and in proteomics for the identification and characterization of proteins in cell lysates, isolated organelles, or purified multisubunit complexes (1–3). Protein separation technologies based on centrifugation, electrophoresis, or chromatographic methods are readily interfaced to MS in an online or off-line fashion. Proteins

are then converted into peptides by treatment with sequence-specific proteases or chemical reagents, since peptides are more amenable to MS and MS/MS analysis than intact proteins. The availability of robust and sensitive matrix-assisted laser desorption/ionization MS (MALDIMS) (4) and electrospray ionization MS (ESI-MS) (5) instruments makes advanced MS technology accessible to molecular cell biologists, biochemists, and proteomics researchers. MS using either MALDI or ESI as the ionization method enables accurate mass determination of peptides. Peptide mass fingerprinting by MALDI-MS and subsequent sequence database searching is widely used for identification of proteins that are isolated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) or by two-dimensional gel electrophoresis (2DE) gels. In most studies it is also desirable to obtain amino acid sequence information, because the accurately determined molecular mass in combination with a partial amino acid sequence of a peptide are very specific probes for protein identification (1,6). Furthermore, mass analysis and amino acid sequencing by MS/MS may reveal the presence of PTMs at individual amino acid residues in proteins.

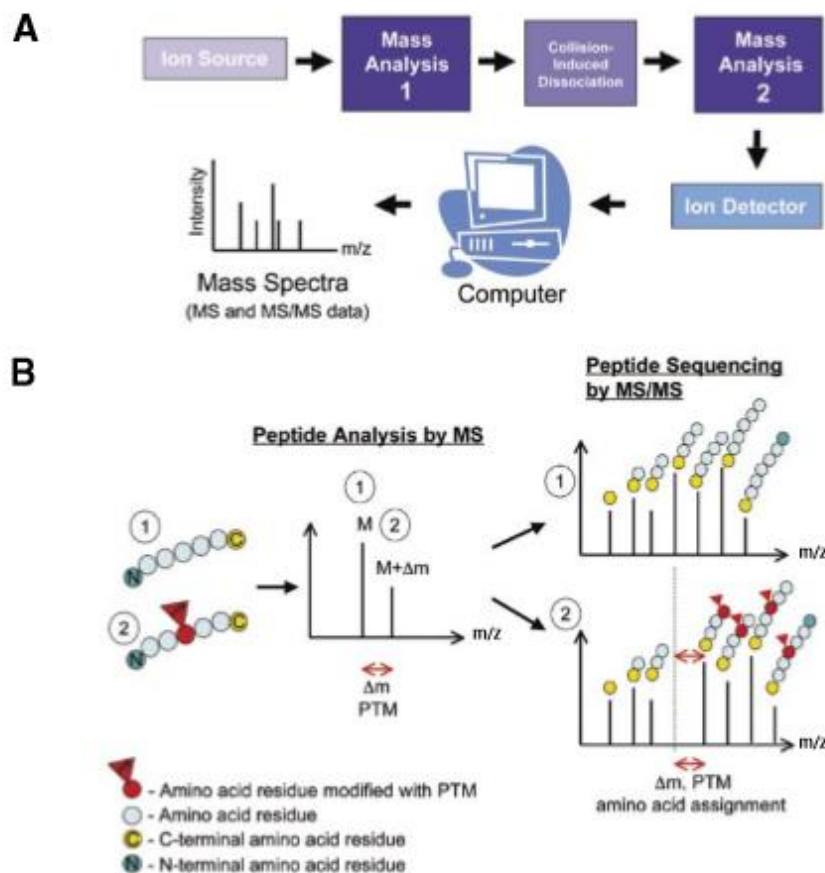


Figure 1. Tandem mass spectrometry (MS/MS) for mapping posttranslational modifications (PTMs). MS/MS is a very sensitive, accurate, and efficient method for sequencing of peptides via the generation and detection of sequence-specific fragment ions. In addition to providing the means to identify proteins, MS also offers tools to interrogate proteins for PTMs. PTM-specific mass increments of peptides and amino acid residues or diagnostic fragment ions in mass spectra reveal the presence of PTMs. (A) Modular setup of MS/MS experiment. A mass spectrometer consists of an ion source, in which the sample molecules are ionized, a mass analyzer, which separates the ions according to their mass-to-charge ratio (m/z), and a detector that records the ions. A sophisticated computer system controls the mass spectrometer and the data-dependent acquisition functions. (B) MS and MS/MS spectra of peptide carrying a PTM. First, the MS spectrum is acquired to determine the molecular mass of the peptides. Next, peptides are in turn selected for MS/MS. Fragmentation of the peptide amide bond produces a set of fragment ions that generate a ladder-like readout of the sequence in the tandem mass spectrum. The presence of a PTM will change the mass of the modified amino acid residue and of the peptide. MS/MS often reveals the mass of the PTM and the identity and position of the modified amino acid residue.

PEPTIDE SEQUENCING BY MS/MS

The tandem mass spectrometer provides the means for amino acid sequencing of peptides. It enables gasphase isolation of individual peptide ion species followed by collision-induced dissociation (CID) and detection of the resultant amino acid sequence specific fragment ions (Figure 1). The MS/MS experiment consists of several stages of mass analysis and ion manipulation. First, the masses of the sample analytes are determined in a MS survey scan (first

MS experiment). Second, the peptide ion of interest is isolated via its mass-to-charge ratio (m/z) value (i.e., by filtering away other ion species that have a different m/z value). Third, the selected peptide ion species is activated (e.g., by collisions with an inert gas such as Argon that imparts internal energy into the ions and thereby induces their fragmentation). Last, the m/z values of the fragment ions are determined (second MS experiment). The most labile bonds in peptides are generally the backbone amide bonds, leading to breakage of the peptide backbone in between amino acids. Hence, tandem mass spectra of peptides contain a series of sequence-revealing fragment ions (8). The fragment ion signals reflect the amino acid sequence as read from either the N-terminal (b-ion series) or the C-terminal (y-ion series) direction (9). The identities of the individual amino acids are revealed by the mass differences between the signals in b-ion series or y-ion series (Figure 1). MS/MS is a sensitive and fast technique for peptide sequencing. Depending on the type of instrumentation and the experimental design, 10–500 ng pure protein are needed to obtain sufficient peptide data for protein identification. Microgram amounts of protein are needed for proteomics analysis of complex protein samples and organelles. Coupling of capillary liquid chromatography to tandem mass spectrometry (LC-MS/MS) allows for automated analysis of complex peptide mixtures by data-dependent acquisition of peptide mass spectra and tandem mass spectra. Typically, 1–5 peptides are analyzed by MS/MS per second. Hence, hundreds or even thousands of peptides can be analyzed in a single LC-MS/MS experiment to reveal the inventory of proteins in biological samples. Large-scale MS experiments in proteomics are facilitated by computational data analysis and database searching algorithms that enable protein identification and quantitation. A series of computational methods for the assignment of peptide sequences based on automated interpretation of MS/MS spectra and protein sequence database searching have been developed (6,10,11). These tools are often interfaced to or integrated with MS software provided by manufacturers to give an efficient data processing pipeline for annotation of peptides and the equivalent proteins.

PTM ANALYSIS BY MS/MS

Accurate determination of the mass increment at the protein or peptide level will aid in defining the type of modification. Intact mass determination of purified proteins is a useful method for determining modification and processing events. Comparison of the experimentally detected intact molecular mass with the calculated mass obtained from the amino acid sequence of the protein will reveal any discrepancies, and the mass difference or deviation may define the

modification. In addition, the mass spectrum will often reveal heterogeneous modification and processing of the protein, resulting in multiple species that each correspond to an individual modification state. A large class of PTMs is represented by chemical moieties that are covalently attached to proteins by various enzymes. Examples are phosphorylation (+80 Da), sulfation (+80 Da), nitration (+45 Da), O-glycosylation (>203 Da), and acylation (>200 Da) (Table 1 and references therein).

Another class of PTM is proteolytic processing that leads to mass deficit. Removal of the leader methionine residue (131 Da) is an example of such a processing event. Processing of precursor polypeptides for removal of signal sequences to generate biologically active proteins is yet other types of processing. Amino acid oxidations are also common PTMs, especially on methionine and tryptophan residues, leading to mass increments of 16 or 32 Da (Table 1). The formation of disulfide bridges between cysteine residues leads to a mass change of -2 Da. Thus, mass analysis of intact proteins or the derived peptides can reveal modifications and also provide information on the type of the PTMs. However, assignment of the PTM to a specific amino acid residue necessitates sequencing of the protein or peptide. MS/MS facilitates mass determination and sequencing of peptides and thereby also the detection of site-specific PTMs (3). The chemical stability of the PTM is decisive for its efficient detection in MS/MS. Certain PTMs will remain intact during MS and MS/MS experiments. For example, acetyl-lysine is a very stable PTM that leads to a mass increment of 42 Da of the intact peptide. Upon MS/MS of a lysineacetylated peptide, all the fragment ions that contain the AcLys residue exhibit a +42 Da mass increment relative to the unmodified peptide (Table 1). Less stable PTMs are phosphoserine and phosphothreonine, which often eliminate phosphoric acid (or a phosphate group and water) during MS/MS (12). Thus, all the fragment ions that contained these phosphoamino acid residues will typically exhibit a mass deficit of 98 Da (-H₃PO₄), also known as a neutral loss, relative to the intact phosphopeptide. These Δm values are very useful for annotation of PTM peptide MS/MS spectra, as they aid in identifying and assigning the modified residues (Table 1). In addition to the amino acid sequence-specific peptide fragment ions, also PTM-specific signals from modified amino acids exist. These PTM-specific fragment ions appear in the lower mass range of tandem mass spectra, typically below m/z 400. Certain amino acids are detected as immonium ion fragments that constitute a fingerprint for the presence of these amino acid residues in a peptide. Several types of modified amino acid residues also

provide diagnostic ion signals. Acetyl-lysine generates ion signals at m/z 126.091 and 143.118 (Table 1 and References 13–15). The presence of these signals in MS/MS spectra is a good indication that the corresponding peptide contains Ac-Lys. Similarly, phosphotyrosine generates a diagnostic signal at m/z 216.042 (16). Glycopeptides generate glycan-specific fragments (oxonium ions) and acylated peptides produce lipid-specific signals that are useful for verification of these types of PTM in MS/MS (Table 1). False positive signals due to interference from regular peptide fragment ions should be considered when the resolution and mass accuracy is insufficient to distinguish PTM-specific signals from internal peptide fragments, such as di- or tri-amino acids.

MS/MS provides a number of useful analytical features that take advantage of diagnostic ion signals and neutral loss products. Precursor ion scanning is a selective and sensitive method for specific detection of only those peptides that carry a PTM that generate a unique diagnostic ion. For example, phosphopeptides generate a strong m/z 79 ion signal in the negative ion MS/MS mode. Thus, precursor ion scans for this ion will identify only phosphopeptides in the sample, whereas regular peptides remain undetected. A number of glycans and lipids generate fragment ions that can be explored for specific detection of PTM peptides (Table 1). The precursor ion scanning method is extremely useful for the analysis of phosphorylation sites and other PTMs in individual, purified proteins (Table 1). Neutral loss analysis is in proteomics typically performed by ion trap instruments that allow very fast MS/MS and multistage mass spectrometry (MS/MS/MS) analysis. The principle is based on the lability of PTMs, such as phosphoserine/threonine, that readily undergoes gas-phase elimination of a PTM-related chemical group during MS/MS. For example, the neutral loss of H_3PO_4 observed upon MS/MS of phosphopeptides will trigger a MS/MS/MS analysis of the product, to reveal the sequence of the PTM peptide and the position of the PTM (Figure 2).

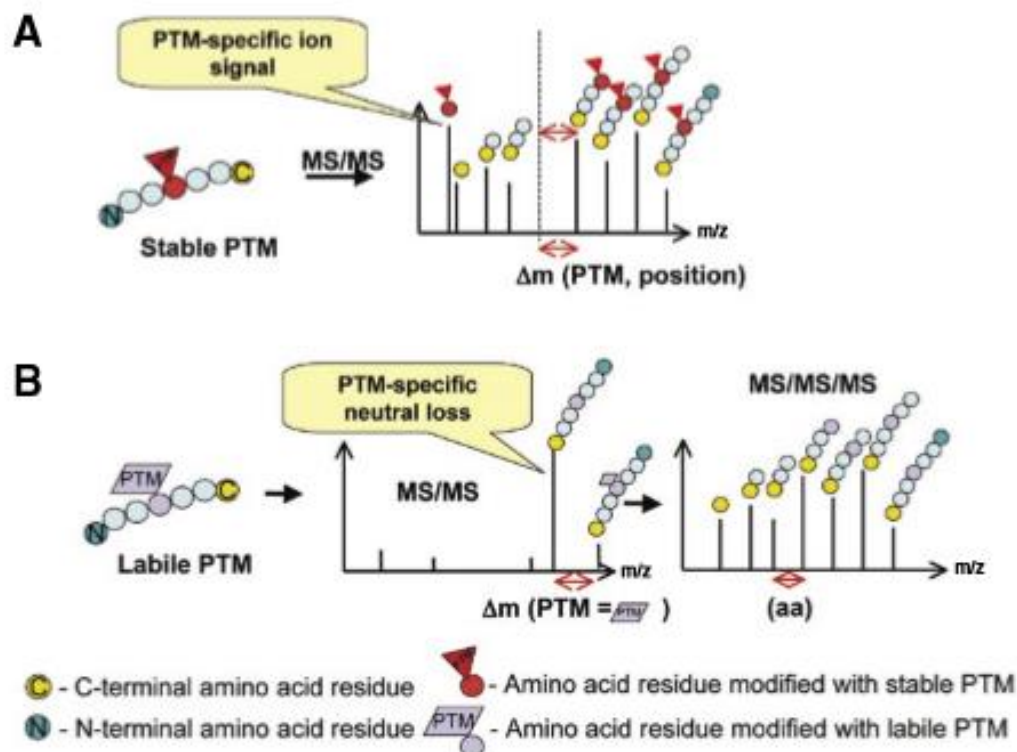


Figure 2. Tandem mass spectrometry (MS/MS) of posttranslational modifications (PTMs). (A) MS/MS of a stable PTM leads to detection of fragment ions that reveal the identity and position of the modified amino acid (aa) residue(s). In addition, diagnostic fragment ions originating from the modified amino acid residue are often observed. Such signals are highly useful when searching for modified peptides in large MS/MS data sets. (B) MS/MS of a labile PTM typically leads to the observation of one strong signal in the MS/MS spectrum and no significant sequence-revealing fragment ion series. Multistage mass spectrometry (MS/MS/MS) is useful to further characterize the fragment ion species by sequencing. This approach is very efficient for sequencing of phosphoserine and phosphothreonine peptides. m/z , mass-to-charge ratio.

COMPUTATIONAL TOOLS FOR PTM ASSIGNMENT IN MS/MS DATA SETS

Several computational methods for automated annotation of PTM in peptides exist. These algorithms analyze the MS and MS/MS data, taking into account the Δm values and sometimes also neutral losses and diagnostic ions for the PTM of interest. Once a set of proteins is identified in a MS/MS-based proteomics experiment, it is possible to perform a more detailed analysis of the retrieved protein sequences and the MS/MS data set. The initial database search is typically performed with only a limited set of specified modifications, including alkylated cysteines and oxidized methionines, in order to minimize search space and to avoid false positive PTM assignments. In a second stage, the molecular masses of the unmodified proteins and the corresponding proteolytic peptides are readily calculated from the protein sequences obtained by the initial sequence database search. As mentioned previously, PTM will lead to a mass

increment or mass deficit (Δm) of the modified peptide relative to the unmodified species. Thus, it becomes feasible to inquire whether any of the initially unassigned MS/MS spectra can be matched to posttranslationally modified peptides, given a list of putative modifications and their Δm values and the list of predicted and observed peptide masses. Although highly useful, computational sequence annotation tools should be used with care, because they may produce false positives in cases where the mass accuracy or signal-to-background level of the MS and MS/MS data are not sufficient to make unambiguous assignments. This is particularly important in modification-specific proteomics, in which protein identification is often based on the detection and sequencing of only a few PTM peptides per protein. The presence of multiple modifications or various different types of modifications in a peptide may also complicate MS and MS/MS data interpretation.

STRATEGIES FOR DETERMINATION OF PTM IN PROTEOMICS

Mapping of PTMs in proteomics is a demanding task because most PTMs are low abundance and/or substoichiometric and some PTMs are labile during MS and MS/MS. In addition, many modifications are hydrophilic, which complicates PTM sample handling and purification prior to MS (27). The presence of PTMs may affect the cleavage efficiency of proteases, such as trypsin, to generate unexpected or large peptide products. Certain PTMs will reduce the ionization and detection efficiency in MS. Multisite PTMs may generate very complicated MS and MS/MS data sets that are difficult to interpret. For these reasons, it is often useful to consider and explore several approaches for mapping of PTMs in proteomics (Figure 3). Generally, it is recommended to reduce the complexity of the sample as much as possible prior to MS analysis for mapping of PTMs. It is not advisable to pursue PTM mapping by direct LC-MS and MS/MS analysis of crude cell lysates, as PTMs are rarely detected in such experiments. This is mainly due to the presence of many abundant unmodified peptides, the inadequate peptide separation by LC, and the limitations of MS/MS for peptide sequencing of very complex mixtures. Biochemical purification of cellular compartments, organelles, protein complexes, or of individual proteins has proved to be a successful first stage for analysis of PTMs, as it dramatically reduces the complexity of the initial protein sample. As an alternative, PTM-specific reagents are useful for selective enrichment of PTM proteins and peptides prior to mass spectrometry. Mapping of PTMs in proteins is achievable by a series of approaches, including (i) treatment with multiple proteases and shotgun sequencing by MS/MS; (ii) enrichment of modified proteins or peptides

prior to MS/MS sequencing; (iii) and PTM-specific MS/ MS and multistage MS. Combinations of these methods have proven successful for comprehensive analysis of PTM in proteomics (Figure 3).

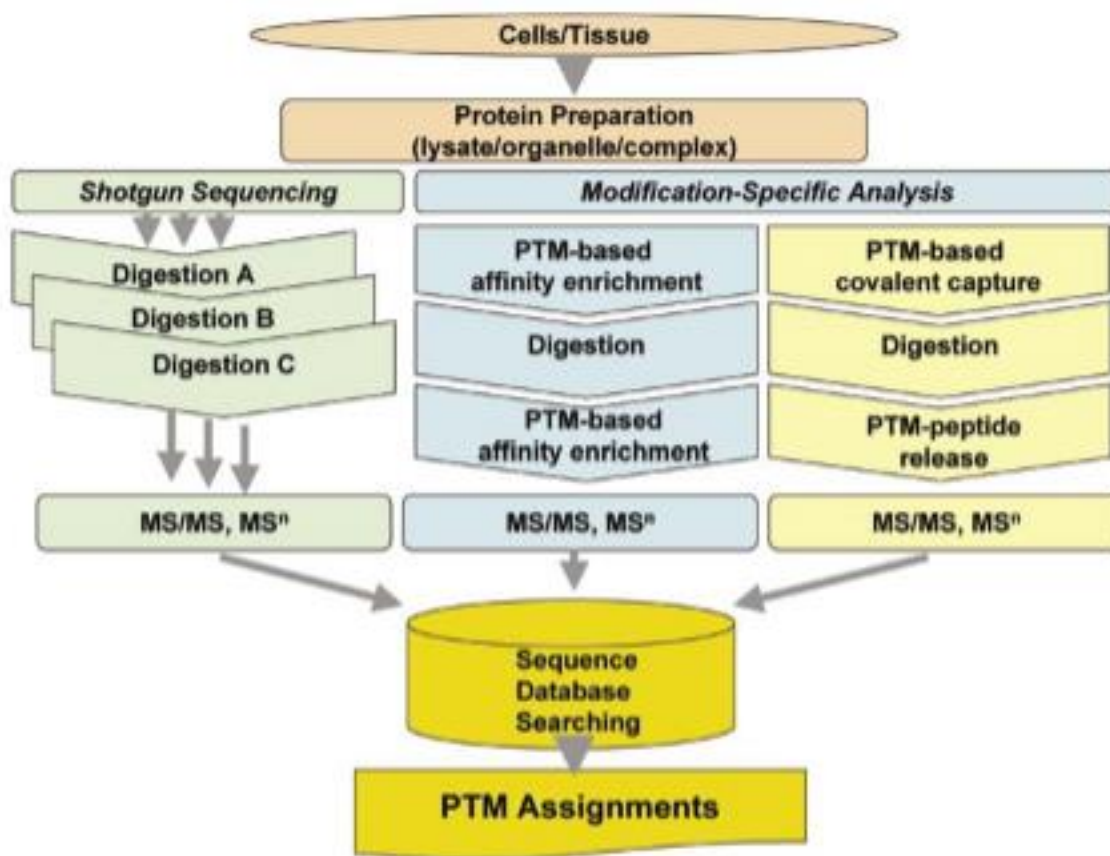


Figure 3. Analytical strategies for mapping posttranslational modifications (PTMs) in proteomics. (Left column) Shotgun sequencing. This method is best suited for subproteomes, simple protein mixtures, and protein complexes, because extensive computational tandem mass spectrometry (MS/MS) data analysis and peptide sequence alignment is required. (Middle column) Modification-specific proteomics based on affinity enrichment of PTM proteins and/or peptides prior to MS analysis. This method is compatible with complex protein mixtures. Several complementary chromatographic peptide separation methods should be used prior to MS/MS analysis. (Right column) Chemistry-based methods for PTM-specific covalent capture of proteins and/or peptides. Often based on β -elimination/Michael addition chemistry, phosphoramidate chemistry, or other PTM-specific chemistries.

Shotgun Sequencing

The shotgun sequencing approach is based on using multiple proteases with different cleavage specificity to generate complementary and redundant sets of overlapping peptides. The peptide mixtures are then analyzed by LC-MS/ MS or by multidimensional separation technology and MS/MS (LC/LC-MS/ MS) for comprehensive sequencing of proteins. Database searching and

computational assembly of the sequence data then provides an overview of the protein sequences and annotation of PTMs and amino acid substitutions (28,29), and it is also amendable to mapping of modifications in purified proteins and protein complexes.

PTM-Specific Enrichment

Large-scale analysis of PTMs is facilitated by PTM-specific protein and peptide enrichment methods, such as PTM-directed affinity chromatography or immunoprecipitation with PTM-specific antibodies. Phosphoproteins can be purified by PTM-specific affinity resins (30) or anti-phosphoamino acid antibodies (31,32). Subsequent digestion of protein and LC-MS/MS analysis of peptides will reveal the identity of the retrieved proteins and sometimes allow site-specific assignment of phosphorylation sites in the recovered proteins. However, it is often also beneficial to enrich for PTM peptides prior to MS analysis in order to improve sensitivity and specificity. Phosphopeptides can be recovered by antibodies (33), immobilized metal affinity chromatography (IMAC) or by TiO₂ columns (37,38) prior to MS/MS analysis. Strong cation exchange and anion exchange chromatography have also proven useful for reducing peptide complexity prior to MS/MS based detection and sequencing of modified peptides, including phosphopeptides (24,39). Using combinations of these methods, large-scale phosphoproteomics have revealed thousands of phosphorylated sites in proteins from various species (24,26,39,40). Glycoproteins can be enriched by using lectins, which are sometimes referred to as sugar-specific antibodies (41,42). Glycoproteins from serum were retrieved and identified by a two-stage lectin enrichment procedure that first targeted the intact glycoproteins and then the proteolytically derived glycopeptides. The glycopeptide sample was then treated with N-glycosidase F in the presence of 18-O water and analyzed by LC-MS/MS. The 18-O labeling enabled site-specific assignment of glycosylation sites by signature ion signals MS/MS (43). An alternative strategy used lectins for glycoprotein enrichment followed by hydrophilic interaction chromatography (HILIC) for glycopeptide enrichment. Glycosidase D/H treatment and MS/MS then facilitated protein identification and assignment of glycosylation sites (44). GPI-anchored proteins, another class of glycoproteins, were identified by using modification-specific enzymes (phospholipases) to selectively release GPI-anchored proteins from plasma membrane preparations in a two phase detergent system, followed by their identification by LC-MS/MS and bioinformatics sequence analysis (45). O-glycosylation is also becoming tractable by using affinity enrichment and advanced MS methods (46). Genetic methods are also useful for

enrichment of PTM proteins. Ubiquitylated and SUMOylated proteins are amenable to purification and identification by integration of genetic tags into the modifying proteins (47).

PTM-Specific Chemistry

Certain modifications are amenable to chemical conversion to stable, tractable species. This is particularly useful for PTMs that are labile during MS and MS/MS analysis. O-phosphoSer/Thr and O-GlcNAc-Ser/Thr readily undergo β -elimination to generate intermediates that are substrates for Michael addition of alkylating groups. Such methods have been used for selective recovery of phosphoproteins and glycoproteins by solid-phase chemistry or by affinity purification prior to protein digestion and identification by MS/MS sequencing (46,48). The introduction of fluoruous affinity tags helps recover modified peptides by taking advantage of the unique chromatographic properties of fluoruous compounds (19). Another option is to use β -elimination/Michael addition reactions to introduce unique mass tags into the modified peptides. The chemical design of the mass tag then provides a diagnostic mass fingerprint for the modified peptides, so they are readily distinguished from regular peptides (49). Other chemistries and resins, including dendrimers, have also been applied for PTM-specific recovery of proteins and peptides (50–52). Tagging-by-substrate methods were applied to identify O-GlcNAc-modified proteins and acylated proteins (53,54).

QUANTIFICATION OF PTM

Dynamic cellular events, such as signal transduction networks, cell cycle progression, and chromatin activity call for quantitative techniques for determination of PTMs. MS-based methods facilitate both absolute and relative quantitation of peptides and their PTM. Absolute quantitation of peptides by MS is achievable by using internal standard peptides for selected proteins and defined PTMs (55). Relative quantitation of PTM peptides is obtained by either peptide intensity profiling (PIP) by LC-MS or by stable isotope labeling (SIL) by using stable isotope-encoded chemical precursor molecules or alkylating reagents (1,3). The functionally important PTMs will exhibit variations in abundance as a result of the perturbation, and they are identified by comparison of the peptide ion signals obtained from two or more defined cellular states, typically a control experiment and one or more perturbed states. The relative abundance of individual peptides is then derived by comparative analysis of LC-MS data using retention time and mass (PIP) or by using only the MS ion intensity of isotopically encoded peptides (SIL).

Protein purification and separation

These methods have been used in quantitative phosphoproteomics in various organisms, from microbes to humans, and also in glycoproteomics.



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

UNIT – V -Fundamentals of Genomics and Proteomics– SBI1309

Transcriptomics

The **transcriptome** is the set of all messenger RNA molecules in one cell or a population of cells. It differs from the exome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to the molecular identities. The term can be applied to the total set of transcripts in a given organism, or to the specific subset of transcripts present in a particular cell type. Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome can vary with external environmental conditions. Because it includes all *mRNA* transcripts in the cell, the transcriptome reflects the genes that are being actively expressed at any given time, with the exception of mRNA degradation phenomena such as transcriptional attenuation.

The study of *transcriptomics*, also referred to as expression profiling, examines the expression level of mRNAs in a given cell population, often using high-throughput techniques based on DNA microarray technology. The use of next-generation sequencing technology to study the transcriptome at the nucleotide level is known as RNA-Seq

Transcriptomics is the study of the transcriptome—the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis. Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cell populations, or in response to different treatments

Transcriptomics or global analysis of gene expression, also called genome-wide expression profiling, is one of the tools that is used to get an understanding of genes and pathways involved in biological processes. The idea underlying this approach is called “guilt by association”, which means that genes showing similarity in expression pattern may be functionally related and under the same genetic control mechanism.

Common technologies for genome-wide or high-throughput analysis of gene expression are cDNA microarrays and oligo-microarrays, cDNA-AFLP and SAGE.

cDNA microarray

A **cDNA microarray** works by using the ability of a given mRNA molecule to bind specifically to, or hybridiseA **cDNA microarray** works by using the ability of a given mRNA molecule to bind specifically to, or hybridise to, its original DNA coding sequence in the form of a cDNA template spotted on an array. cDNA microarray experiments typically

involve hybridising two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorescent dye (usually a red fluorescent dye, Cyanine 5 (Cy5) and a green-fluorescent dye, Cyanine 3 (Cy3)), on a single glass slide that has been spotted with (several thousands of) cDNA probes. Because of competitive binding between the two samples, the ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two samples. Thus data from cDNA microarrays only provide information on the relative expression of the genes. Common features in the layout of a cDNA microarray slide are shown in the figure below.

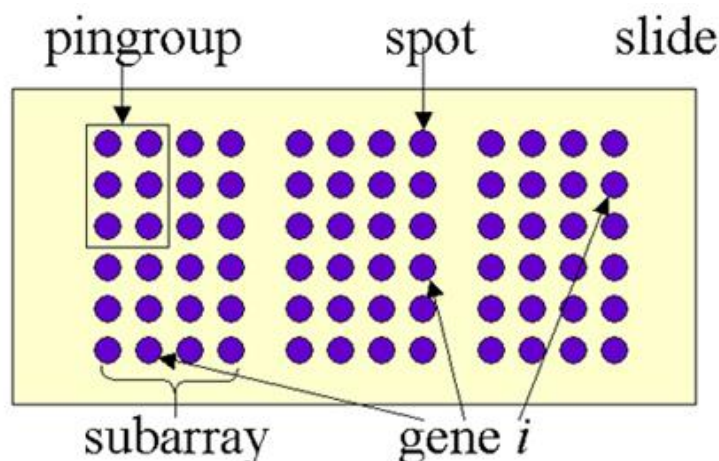


Figure: Common features in the layout of a cDNA microarray slide. The fundamental units on which measurements are obtained are spots containing cDNA clones fixed to a glass substrate. Robotic printing devices used to generate the spots often work with multiple printing tips or pins. Printing may also be done in blocks (subarrays). The same clone (gene) may be printed multiple times on a single slide.

Oligonucleotide arrays

Besides using cDNA clones as probes on an array, oligonucleotides of around 20 nucleotides can also be used as probe (for example GeneChip® oligonucleotide arrays from Affymetrix). High-density **oligonucleotide arrays** provide direct information about the expression levels in an mRNA, because these microarrays usually are hybridised with only one sample. Instead of one or several cDNA clones per gene, an oligo-microarray contains two times a set of probe pairs for each gene (see figure below). One set consists of ‘perfect match’ oligonucleotides that are designed on non-conserved regions in a gene. The other sets contains ‘mismatch oligonucleotides’ that are identical to the set of ‘perfect match’ oligonucleotides except for the nucleotide in the middle of the sequence. In the presence of a specific RNA in the hybridisation solution, the perfect match probes will hybridise more

strongly on average than their mismatch partners. This assumption is used to determine the presence or absence of this specific RNA. For each probeset, the value that is usually taken as representative for the expression level of the corresponding gene (the quantitative RNA abundance), is the average difference between the set of perfect match probes minus the set of mismatch probes. In addition, the mismatched oligonucleotides are used to calculate cross-hybridisation and local background signals.

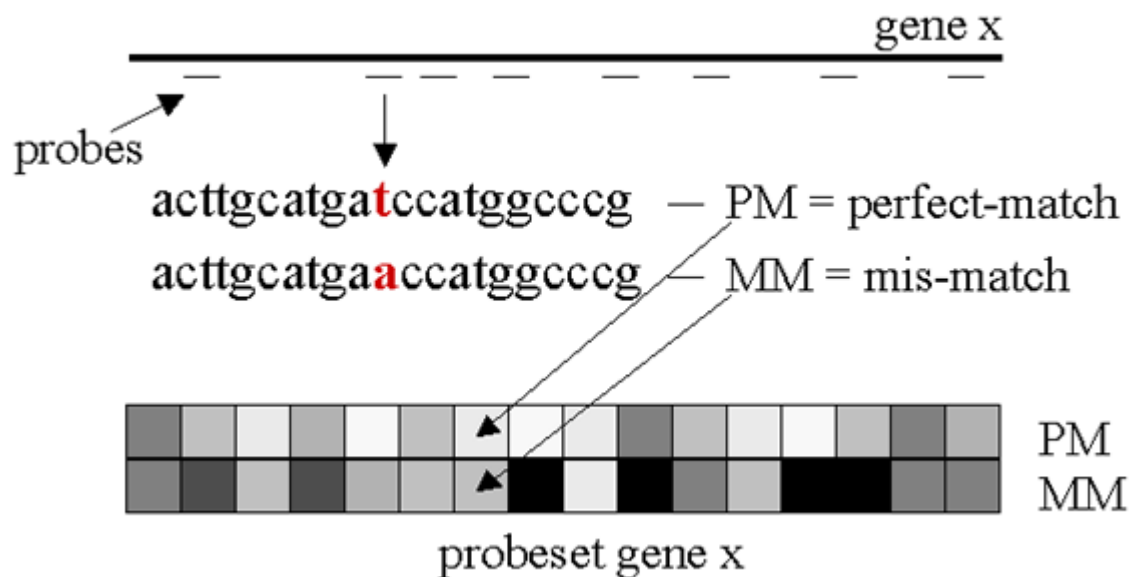


Figure: The principle of oligonucleotide-microarrays. The probes correspond to short oligonucleotide sequences thought to be representative for a given gene. Each oligonucleotide is represented by two probes: one with the exact sequence and one with a mismatch nucleotide. Hybridisation occurs at a critical temperature where a single nucleotide mismatch can prevent hybridisation. White = strong hybridisation, black = no hybridisation. Based upon the hybridisation ratio between PM and MM, it is determined whether a gene is 'Absent', 'Marginal' or 'Present'. For gene x there is quite a lot of hybridisation to the MM probes compared to the PM probes and therefore the analysis software will probably report the presence of this gene-transcript as 'Marginal'.

Advantages and disadvantages

Microarray experiments allow for comparison of gene expression profiles between two mRNA samples (e.g. treatment vs control, or treatment 1 vs treatment 2). However, the most important advantage of microarray-based technology is that large data sets from different experiments can be combined together in a single database, which allows gene

expression profiles from either different samples or samples obtained using different treatments to be compared with each other and analysed together (e.g. time series). In medical biology, microarrays are most often used for class prediction. Which means using the gene expression profile of a tissue sample to assign a particular patient to an already defined class. The class prediction is based on marker genes that are specifically expressed during a certain disease or treatment. It is envisaged that global surveys of gene expression will identify additional marker genes that may be used to group patients into molecularly relevant categories.

Limitations:

- Transcript profiling produces extremely large data sets, and even relatively simple studies can run into thousands of data points. It is self-evident that such data sets cannot be organized on simple spreadsheets, but instead requires effective database resources for management and analysis.
- With cDNA microarrays it is difficult to distinguish among different transcripts from genes belonging to the same gene family. In order to prevent cross-hybridisation long oligonucleotides may be used instead of cDNA.
- Oligo-arrays: oligonucleotide design can avoid cross-hybridisation, but it is very expensive and requires accurate gene annotation. Pre-made oligonucleotide arrays can be bought, which is less costly, but also less flexible than cDNA arrays, because the experimenter cannot select the probes.
- Oligo-arrays use smaller quantities of RNA. The hybridisation signal is linearly related up to 500-fold mRNA abundance compared to 10-fold in cDNA microarrays. Oligo-arrays have a higher detection specificity (1 out of 10⁶), but are usually only hybridised with one mRNA sample.
- Transcript profiling using microarrays is limited to the genes that are represented on the chip. Only those genes for which either the DNA sequence or a cDNA clone is available can be used in microarray studies. However, probably not all genes are known yet, because there are transcripts whose corresponding genes were not, or wrongly, identified during the genome annotation procedure, and there are genes that are thought to produce no transcript. These genes probably will not be represented on a microarray.
- When cDNA libraries are used to produce the probes, rarely expressed genes are often missing, because abundant messengers are over-represented. For this reason, many important regulatory genes will be overlooked.

cDNA-AFLP: cDNA-amplified fragment length polymorphism analysis

cDNA-AFLP is an improvement of traditional differential display techniques. It is a PCR-based method which starts with cDNA synthesis from total or mRNA using random hexamers as primers. The obtained fragments are digested with two restriction enzymes, normally a 4-cutter and a 6-cutter, and adapters are ligated to the ends of the fragments. In the first amplification step only those fragments are amplified that were digested by both restriction enzymes and thus have different adapters at the end. In the following amplification steps, the complex starting mixture of cDNA is fractionated into smaller subsets by selective PCR amplification using primers on the adapters that contain one or more extra nucleotides. By increasing the stringency of the PCR amplification (adding more additional nucleotides to the primers), the sensitivity of the analysis can be increased. In this way, also genes with a low expression level can be detected.

The fragments that are amplified are roughly 100-400 bp. These fragments are separated on high-resolution gels. The differences in the intensity of the bands that can be observed provide a good measure of the relative differences in the levels of gene expression. Further characterisation of interesting transcripts often requires the identification of the corresponding full-length cDNA.

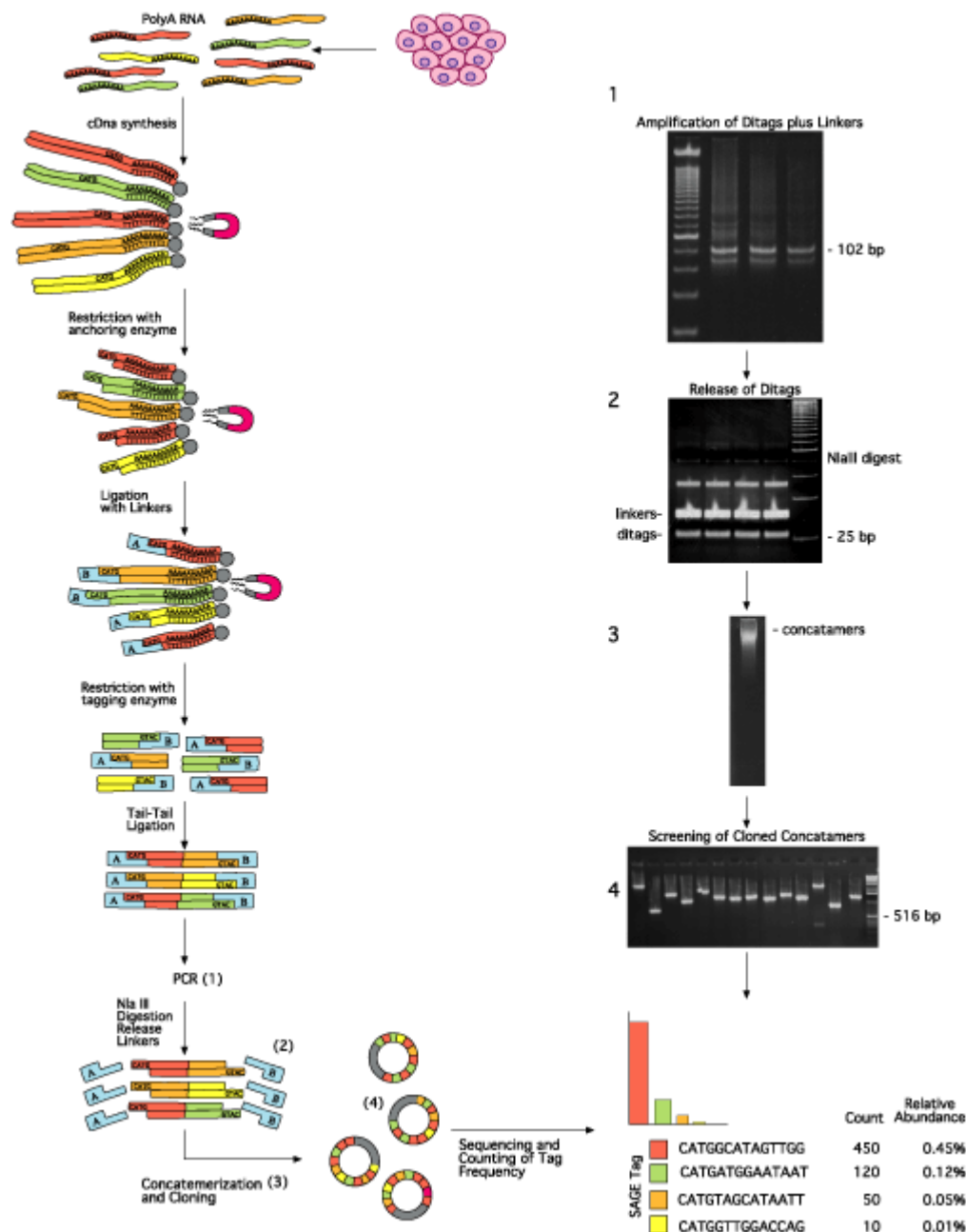
cDNA-AFLP can generate a global overview of gene expression, but it involves a great amount of PCR reactions. In addition, separately obtained data sets cannot readily be compared, which is in contrast to SAGE and microarrays data. However, using cDNA-AFLP, accurate gene expression profiles can be determined by quantitative analysis of band intensities. Furthermore, the sensitivity and specificity of the method allows the detection of poorly expressed genes and the determination of subtle differences in transcriptional activity.

Serial analysis of gene expression

Serial analysis of gene expression (SAGE) is a technique used by molecular biologists to produce a snapshot of the messenger RNA population in a sample of interest in the form of small tags that correspond to fragments of those transcripts. The original technique was developed by Dr. Victor Velculescu at the Oncology Center of Johns Hopkins University and published in 1995.[1] Several variants have been developed since, most notably a more robust version, LongSAGE, RL-SAGE and the most recent Super SAGE. Many of these have improved the technique with the capture of longer tags, enabling more confident identification of a source gene

SAGE experiments proceed as follows:

- The mRNA of an input sample (e.g. a tumour) is isolated and a reverse transcriptase and biotinylated primers are used to synthesize cDNA from mRNA.
- The cDNA is bound to Streptavidin beads via interaction with the biotin attached to the primers, and is then cleaved using a restriction endonuclease called an anchoring enzyme (AE). The location of the cleavage site and thus the length of the remaining cDNA bound to the bead will vary for each individual cDNA (mRNA).
- The cleaved cDNA downstream from the cleavage site is then discarded, and the remaining immobile cDNA fragments upstream from cleavage sites are divided in half and exposed to one of two adapter oligonucleotides (A or B) containing several components in the following order upstream from the attachment site: 1) Sticky ends with the AE cut site to allow for attachment to cleaved cDNA; 2) A recognition site for a restriction endonuclease known as the tagging enzyme (TE), which cuts about 15 nucleotides downstream of its recognition site (within the original cDNA/mRNA sequence); 3) A short primer sequence unique to either adapter A or B, which will later be used for further amplification via PCR.
- After adapter ligation, cDNA are cleaved using TE to remove them from the beads, leaving only a short "tag" of about 11 nucleotides of original cDNA (15 nucleotides minus the 4 corresponding to the AE recognition site).
- The cleaved cDNA tags are then repaired with DNA polymerase to produce blunt end cDNA fragments.
- These cDNA tag fragments (with adapter primers and AE and TE recognition sites attached) are ligated, sandwiching the two tag sequences together, and flanking adapters A and B at either end. These new constructs, called ditags, are then PCR amplified using anchor A and B specific primers.
- The ditags are then cleaved using the original AE, and allowed to link together with other ditags, which will be ligated to create a cDNA concatemer with each ditag being separated by the AE recognition site.
- These concatemers are then transformed into bacteria for amplification through bacterial replication.
- The cDNA concatemers can then be isolated and sequenced using modern high-throughput DNA sequencers, and these sequences can be analysed with computer programs which quantify the recurrence of individual tags.



The output of SAGE is a list of short sequence tags and the number of times it is observed. Using sequence databases a researcher can usually determine, with some confidence, from which original mRNA (and therefore which gene) the tag was extracted.

Statistical methods can be applied to tag and count lists from different samples in order to determine which genes are more highly expressed. For example, a normal tissuesample can be compared against a corresponding tumour to determine which genes tend to be more (or less) active.

Although SAGE was originally conceived for use in cancer studies, it has been successfully used to describe the transcriptome of other diseases and in a wide variety of organisms.

Comparison to DNA microarrays

The general goal of the technique is similar to the DNA microarray. However, SAGE sampling is based on sequencing mRNA output, not on hybridization of mRNA output to probes, so transcription levels are measured more quantitatively than by microarray. In addition, the mRNA sequences do not need to be known a priori, so genes or gene variants which are not known can be discovered. Microarray experiments are much cheaper to perform, so large-scale studies do not typically use SAGE. Quantifying gene expressions is more exact in SAGE because it involves directly counting the number of transcripts whereas spot intensities in microarrays fall in non-discrete gradients and are prone to background noise

Variant Protocols: miRNA cloning

MicroRNAs, or miRNAs for short, are small (~22nt) segments of RNA which have been found to play a crucial role in gene regulation. One of the most commonly used methods for cloning and identifying miRNAs within a cell or tissue was developed in the Bartel Lab and published in a paper by Lau et al. (2001). Since then, several variant protocols have arisen, but most have the same basic format. The procedure is quite similar to SAGE: The small RNA are isolated, then linkers are added to each, and the RNA is converted to cDNA by RT-PCR. Following this, the linkers, containing internal restriction sites, are digested with the appropriate restriction enzyme and the sticky ends are ligated together into concatamers. Following concatenation, the fragments are ligated into plasmids and are used to transform bacteria to generate many copies of the plasmid containing the inserts. Those may then be sequenced to identify the miRNA present, as well as analysing expression levels of a given miRNA by counting the number of times it is present, similar to SAGE.

Pharmacogenomics

Pharmacogenomics (a portmanteau of pharmacology and genomics) is the study of the role of genetics in drug response. It deals with the influence of acquired and inherited genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with drug absorption, distribution, metabolism and elimination, as well as drug receptor target effects. The term pharmacogenomics is often used interchangeably with pharmacogenetics. Although both terms relate to drug response based on genetic influences, pharmacogenetics focuses on single drug-gene interactions, while pharmacogenomics encompasses a more genome-wide association approach, incorporating genomics and epigenetics while dealing with the effects of multiple genes on drug response.

Pharmacogenomics aims to develop rational means to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects.[7] Through the utilization of pharmacogenomics, it is hoped that drug treatments can deviate from what is dubbed as the "one-dose-fits-all" approach. It attempts to eliminate the trial-and-error method of prescribing, allowing physicians to take into consideration their patient's genes, the functionality of these genes, and how this may affect the efficacy of the patient's current and/or future treatments (and where applicable, provide an explanation for the failure of past treatments). Such approaches promise the advent of "personalized medicine"; in which drugs and drug combinations are optimized for each individual's unique genetic makeup. Whether used to explain a patient's response or lack thereof to a treatment, or act as a predictive tool, it hopes to achieve better treatment outcomes, greater efficacy, minimization of the occurrence of drug toxicities and adverse drug reactions (ADRs). For patients who have lack of therapeutic response to a treatment, alternative therapies can be prescribed that would best suit their requirements. In order to provide pharmacogenomic-based recommendations for a given drug, two possible types of input can be used: genotyping or exome or whole genome sequencing.[10] Sequencing provides many more data points, including detection of mutations that prematurely terminate the synthesized protein (early stop codon).

Pharmacogenomics is the branch of pharmacology which deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity. By doing so, pharmacogenomics aims to develop rational means to optimise drug therapy, with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects. Such approaches promise the advent

of "personalized medicine", in which drugs and drug combinations are optimised for each individual's unique genetic makeup.

Pharmacogenomics is the whole genome application of pharmacogenetics, which examines the single gene interactions with drugs.

Pharmacogenomics is the study of how an individual's genetic inheritance affects the body's response to drugs. The term comes from the words pharmacology and genomics and is thus the intersection of pharmaceuticals and genetics.

Pharmacogenomics holds the promise that drugs might one day be tailor-made for individuals and adapted to each person's own genetic makeup. Environment, diet, age, lifestyle, and state of health all can influence a person's response to medicines, but understanding an individual's genetic makeup is thought to be the key to creating personalized drugs with greater efficacy and safety. Pharmacogenomics combines traditional pharmaceutical sciences such as biochemistry with annotated knowledge of genes, proteins, and single nucleotide polymorphisms.

Pharmacogenomics was first recognized by Pythagoras around 510 BC when he made a connection between the dangers of fava bean ingestion with hemolytic anemia and oxidative stress. Interestingly, this identification was later validated and attributed to deficiency of G6PD in the 1950s and called favism.^{[11][12]} Although the first official publication dates back to 1961,^[13] circa 1950s marked the unofficial beginnings of this science. Reports of prolonged paralysis and fatal reactions linked to genetic variants in patients who lacked butyryl-cholinesterase ('pseudocholinesterase') following administration of succinylcholine injection during anesthesia were first reported in 1956.^{[1][14]} The term pharmacogenetic was first coined in 1959 by Friedrich Vogel of Heidelberg, Germany (although some papers suggest it was 1957). In the late 1960s, twin studies supported the inference of genetic involvement in drug metabolism, with identical twins sharing remarkable similarities to drug response compared to fraternity twins.^[15] The term pharmacogenomics first began appearing around the 1990s.^[11]

The first FDA approval of a pharmacogenetic test was in 2005^[16] (for alleles in CYP2D6 and CYP2C19).

Adverse Drug Reaction.

These three simple words convey little of the horror of a severe negative reaction to a prescribed drug. But such negative reactions can nonetheless occur. A 1998 study of hospitalized patients published in the *Journal of the American Medical Association* reported

that in 1994, adverse drug reactions accounted for more than 2.2 million serious cases and over 100,000 deaths, making **adverse drug reactions (ADRs)** one of the leading causes of hospitalization and death in the United States. Currently, there is no simple way to determine whether people will respond well, badly, or not at all to a medication; therefore, pharmaceutical companies are limited to developing drugs using a "one size fits all" system. This system allows for the development of drugs to which the "average" patient will respond. But, as the statistics above show, one size does NOT fit all, sometimes with devastating results. What is needed is a way to solve the problem of ADRs before they happen. The solution is in sight though, and it is called pharmacogenomics.

What are the anticipated benefits of pharmacogenomics?

- **More Powerful Medicines**

Pharmaceutical companies will be able to create drugs based on the proteins, enzymes, and RNA molecules associated with genes and diseases. This will facilitate drug discovery and allow drug makers to produce a therapy more targeted to specific diseases. This accuracy not only will maximize therapeutic effects but also decrease damage to nearby healthy cells.

- **Better, Safer Drugs the First Time**

Instead of the standard trial-and-error method of matching patients with the right drugs, doctors will be able to analyze a patient's genetic profile and prescribe the best available drug therapy from the beginning. Not only will this take the guesswork out of finding the right drug, it will speed recovery time and increase safety as the likelihood of adverse reactions is eliminated. Pharmacogenomics has the potential to dramatically reduce the the estimated 100,000 deaths and 2 million hospitalizations that occur each year in the United States as the result of adverse drug response (1).

- **More Accurate Methods of Determining Appropriate Drug Dosages**

Current methods of basing dosages on weight and age will be replaced with dosages based on a person's genetics --how well the body processes the medicine and the time it takes to metabolize it. This will maximize the therapy's value and decrease the likelihood of overdose.

- **Advanced Screening for Disease**

Knowing one's genetic code will allow a person to make adequate lifestyle and environmental changes at an early age so as to avoid or lessen the severity of a genetic disease. Likewise, advance knowledge of a particular disease susceptibility will allow

careful monitoring, and treatments can be introduced at the most appropriate stage to maximize their therapy.

- **Better Vaccines**

Vaccines made of genetic material, either DNA or RNA, promise all the benefits of existing vaccines without all the risks. They will activate the immune system but will be unable to cause infections. They will be inexpensive, stable, easy to store, and capable of being engineered to carry several strains of a pathogen at once.

- **Improvements in the Drug Discovery and Approval Process**

Pharmaceutical companies will be able to discover potential therapies more easily using genome targets. Previously failed drug candidates may be revived as they are matched with the niche population they serve. The drug approval process should be facilitated as trials are targeted for specific genetic population groups --providing greater degrees of success. The cost and risk of clinical trials will be reduced by targeting only those persons capable of responding to a drug.

- **Decrease in the Overall Cost of Health Care**

Decreases in the number of adverse drug reactions, the number of failed drug trials, the time it takes to get a drug approved, the length of time patients are on medication, the number of medications patients must take to find an effective therapy, the effects of a disease on the body (through early detection), and an increase in the range of possible drug targets will promote a net decrease in the cost of health care.

Is pharmacogenomics in use today?

To a limited degree. The cytochrome P450 (CYP) family of liver enzymes is responsible for breaking down more than 30 different classes of drugs. DNA variations in genes that code for these enzymes can influence their ability to metabolize certain drugs. Less active or inactive forms of CYP enzymes that are unable to break down and efficiently eliminate drugs from the body can cause drug overdose in patients. Today, clinical trials researchers use genetic tests for variations in cytochrome P450 genes to screen and monitor patients. In addition, many pharmaceutical companies screen their chemical compounds to see how well they are broken down by variant forms of CYP enzymes (2).

Another enzyme called TPMT (thiopurine methyltransferase) plays an important role in the chemotherapy treatment of a common childhood leukemia by breaking down a class of therapeutic compounds called thiopurines. A small percentage of Caucasians have genetic variants that prevent them from producing an active form of this protein. As a result, thiopurines elevate to toxic levels in the patient because the inactive form of TPMT is unable

to break down the drug. Today, doctors can use a genetic test to screen patients for this deficiency, and the TMPT activity is monitored to determine appropriate thiopurine dosage levels (3).

Drug-metabolizing enzymes

There are several known genes which are largely responsible for variances in drug metabolism and response. The focus of this article will remain on the genes that are more widely accepted and utilized clinically for brevity.

- Cytochrome P450s
- VKORC1
- TPMT

Cytochrome P450

The most prevalent drug-metabolizing enzymes (DME) are the Cytochrome P450 (CYP) enzymes. The term Cytochrome P450 was coined by Omura and Sato in 1962 to describe the membrane-bound, heme-containing protein characterized by 450 nm spectral peak when complexed with carbon monoxide.^[17] The human CYP family consists of 57 genes, with 18 families and 44 subfamilies. CYP proteins are conveniently arranged into these families and subfamilies on the basis of similarities identified between the amino acid sequences. Enzymes that share 35-40% identity are assigned to the same family by an Arabic numeral, and those that share 55-70% make up a particular subfamily with a designated letter.^[18] For example, CYP2D6 refers to family 2, subfamily D, and gene number 6.

From a clinical perspective, the most commonly tested CYPs include: CYP2D6, CYP2C19, CYP2C9, CYP3A4 and CYP3A5. These genes account for the metabolism of approximately 80-90% of currently available prescription drugs.^{[19][20]} The table below provides a summary for some of the medications that take these pathways.

Drug Metabolism of Major CYPs ^{[21][22]}		
Enzyme	Fraction of drug metabolism (%)	Example Drugs
CYP2C9	10	Tolbutamide, ibuprofen, mefenamic acid, tetrahydrocannabinol, losartan, diclofenac
CYP2C19	5	S-mephenytoin, amitriptyline, diazepam, omeprazole, proguanil, hexobarbital, propranolol, imipramine
CYP2D6	20-30	Debrisoquine, metoprolol, sparteine, propranolol,

		encainide, codeine, dextromethorphan, clozapine, desipramine, haloperidol, amitriptyline, imipramine
CYP3A4	40-45	Erythromycin, ethinyl estradiol, nifedipine, triazolam, cyclosporine, amitriptyline, imipramine
CYP3A5	<1	Erythromycin, ethinyl estradiol, nifedipine, triazolam, cyclosporine, amitriptyline, aldosterone

CYP2D6

Also known as debrisoquine hydroxylase (named after the drug that led to its discovery), CYP2D6 is the most well-known and extensively studied CYP gene.^[23] It is a gene of great interest also due to its highly polymorphic nature, and involvement in a high number of medication metabolisms (both as a major and minor pathway). More than 100 CYP2D6 genetic variants have been identified.^[22]

CYP2C19

Discovered in the early 1980s, CYP2C19 is the second most extensively studied and well understood gene in pharmacogenomics.^[21] Over 28 genetic variants have been identified for CYP2C19,^[24] of which affects the metabolism of several classes of drugs, such as antidepressants and proton pump inhibitors.^[25]

CYP2C9

CYP2C9 constitutes the majority of the CYP2C subfamily, representing approximately 20% of the liver content. It is involved in the metabolism of approximately 10% of all drugs, which include medications with narrow therapeutic windows such as warfarin and tolbutamide.^{[25][26]} There are approximately 57 genetic variants associated with CYP2C9.^[24]

CYP3A4 and CYP3A5

The CYP3A family is the most abundantly found in the liver, with CYP3A4 accounting for 29% of the liver content.^[21] These enzymes also cover between 40-50% of the current prescription drugs, with the CYP3A4 accounting for 40-45% of these medications.^[12] CYP3A5 has over 11 genetic variants identified at the time of this publication.^[24]

VKORC1

The vitamin K epoxide reductase complex subunit 1 (VKORC1) is responsible for the pharmacodynamics of warfarin.^[27] VKORC1 along with CYP2C9 are useful for identifying

the risk of bleeding during warfarin administration. Warfarin works by inhibiting VKOR, which is encoded by the VKORC1 gene. Individuals with polymorphism in this have an affected response to warfarin treatment.^[28]

TPMT

Thiopurine methyltransferase (TPMT) catalyzes the S-methylation of thiopurines, thereby regulating the balance between cytotoxic thioguanine nucleotide and inactive metabolites in hematopoietic cells.^[29] TPMT is highly involved in 6-MP metabolism and TPMT activity and TPMT genotype is known to affect the risk of toxicity. Excessive levels of 6-MP can cause myelosuppression and myelotoxicity.^[30]

Codeine, clopidogrel, tamoxifen, and warfarin are a few examples of medications that follow the above metabolic pathways.

Pharmacogenomics is a developing research field that is still in its infancy. Several of the following barriers will have to be overcome before many pharmacogenomics benefits can be realized.

- **Complexity of finding gene variations that affect drug response** - Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence is altered. SNPs occur every 100 to 300 bases along the 3-billion-base human genome, therefore millions of SNPs must be identified and analyzed to determine their involvement (if any) in drug response. Further complicating the process is our limited knowledge of which genes are involved with each drug response. Since many genes are likely to influence responses, obtaining the big picture on the impact of gene variations is highly time-consuming and complicated.
- **Limited drug alternatives** - Only one or two approved drugs may be available for treatment of a particular condition. If patients have gene variations that prevent them using these drugs, they may be left without any alternatives for treatment.
- **Disincentives for drug companies to make multiple pharmacogenomic products** - Most pharmaceutical companies have been successful with their "one size fits all" approach to drug development. Since it costs hundreds of millions of dollars to bring a drug to market, will these companies be willing to develop alternative drugs that serve only a small portion of the population?
- **Educating healthcare providers** - Introducing multiple pharmacogenomic products to treat the same condition for different population subsets undoubtedly will

complicate the process of prescribing and dispensing drugs. Physicians must execute an extra diagnostic step to determine which drug is best suited to each patient. To interpret the diagnostic accurately and recommend the best course of treatment for each patient, all prescribing physicians, regardless of specialty, will need a better understanding of genetics.

Is there a difference between pharmacogenomics and pharmacogenetics?

- **Pharmacogenomics** refers to the general study of all of the many different genes that determine drug behavior.
- **Pharmacogenetics** refers to the study of inherited differences (variation) in drug metabolism and response.

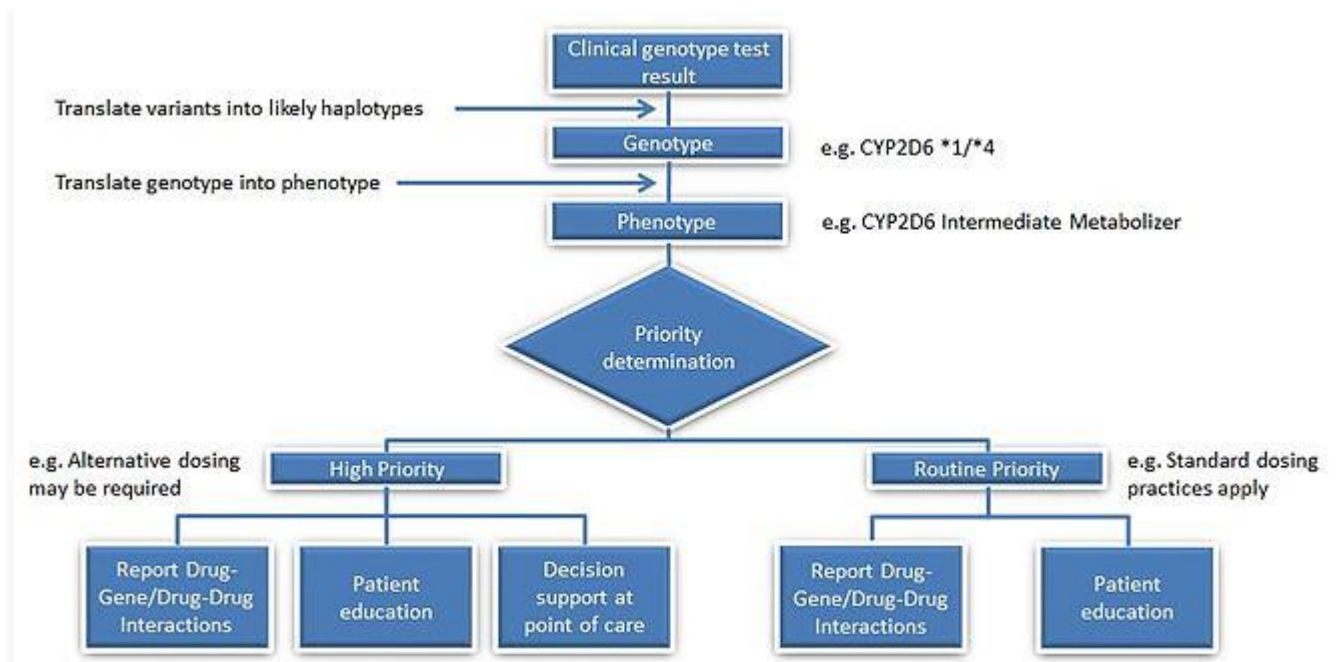
The distinction between the two terms is considered arbitrary, however, and now the two terms are used interchangeably.

Predictive prescribing

Patient genotypes are usually categorized into the following predicted phenotypes:

- Ultra-Rapid Metabolizer: Patients with substantially increased metabolic activity.
- Extensive Metabolizer: Normal metabolic activity;
- Intermediate Metabolizer: Patients with reduced metabolic activity; and
- Poor Metabolizer: Patients with little to no functional metabolic activity.

The two extremes of this spectrum are the Poor Metabolizers and Ultra-Rapid Metabolizers. Efficacy of a medication is not only based on the above metabolic statuses, but also the type of drug consumed. Drugs can be classified into two main groups: active drugs and pro-drugs. Active drugs refer to drugs that are inactivated during metabolism, and Pro-Drugs are inactive until they are metabolized.



An overall process of how pharmacogenomics functions in a clinical practice. From the raw genotype results, this is then translated to the physical trait, the phenotype. Based on these observations, optimal dosing is evaluated.^[29]

For example, we have two patients who are taking codeine for pain relief. Codeine is a pro-drug, so it requires conversion from its inactive form to its active form. The active form of codeine is morphine, which provides the therapeutic effect of pain relief. If person A receives one *1 allele each from mother and father to code for the CYP2D6 gene, then that person is considered to have an extensive metabolizer (EM) phenotype, as allele *1 is considered to have a normal-function (this would be represented as CYP2D6 *1/*1). If person B on the other hand had received one *1 allele from the mother and a *4 allele from the father, that individual would be an Intermediate Metabolizer (IM) (the genotype would be CYP2D6 *1/*4). Although both individuals are taking the same dose of codeine, person B could potentially lack the therapeutic benefits of codeine due to the decreased conversion rate of codeine to its active counterpart morphine.

Each phenotype is based upon the allelic variation within the individual genotype. However, several genetic events can influence a same phenotypic trait, and establishing genotype-to-phenotype relationships can thus be far from consensual with many enzymatic patterns. For instance, the influence of the CYP2D6*1/*4 allelic variant on the clinical outcome in patients treated with Tamoxifen remains debated today. In oncology, genes coding for DPD, UGT1A1, TPMT, CDA involved in the pharmacokinetics of 5-FU/capecitabine, irinotecan, 6-mercaptopurine and gemcitabine/cytarabine, respectively,

have all been described as being highly polymorphic. A strong body of evidence suggests that patients affected by these genetic polymorphisms will experience severe/lethal toxicities upon drug intake, and that pre-therapeutic screening does help to reduce the risk of treatment-related toxicities through adaptive dosing strategies.^[31]

Identification of the genetic basis for polymorphic expression of a gene is done through intronic or exomic SNPs which abolishes the need for different mechanisms for explaining the variability in drug metabolism. The SNPs based variations in membrane receptors lead to multidrug resistance (MDR) and the drug–drug interactions. Even drug induced toxicity and many adverse effects can be explained by genome-wide association studies (GWAS).^[32]

Applications

The list below provides a few more commonly known applications of pharmacogenomics:^[33]

- Improve drug safety, and reduce ADRs;
- Tailor treatments to meet patients' unique genetic pre-disposition, identifying optimal dosing;
- Improve drug discovery targeted to human disease; and
- Improve proof of principle for efficacy trials.

Pharmacogenomics may be applied to several areas of medicine, including Pain Management, Cardiology, Oncology, and Psychiatry. A place may also exist in Forensic Pathology, in which pharmacogenomics can be used to determine the cause of death in drug-related deaths where no findings emerge using autopsy.^[34]

In cancer treatment, pharmacogenomics tests are used to identify which patients are most likely to respond to certain cancer drugs. In behavioral health, pharmacogenomic tests provide tools for physicians and care givers to better manage medication selection and side effect amelioration. Pharmacogenomics is also known as companion diagnostics, meaning tests being bundled with drugs. Examples include KRAS test with cetuximab and EGFR test with gefitinib. Beside efficacy, germline pharmacogenetics can help to identify patients likely to undergo severe toxicities when given cytotoxics showing impaired detoxification in relation with genetic polymorphism, such as canonical 5-FU.^[35]

In cardiovascular disorders, the main concern is response to drugs including warfarin, clopidogrel, beta blockers, and statins

Comparative genomics

Comparative genomics is a field of biological research in which the genomic features of different organisms are compared. The genomic features may include the DNA

sequence, genes, gene order, regulatory sequences, and other genomic structural landmarks. In this branch of genomics, whole or large parts of genomes resulting from genome projects are compared to study basic biological similarities and differences as well as evolutionary relationships between organisms. The major principle of comparative genomics is that common features of two organisms will often be encoded within the DNA that is evolutionarily conserved between them. Therefore, comparative genomic approaches start with making some form of alignment of genome sequences and looking for orthologous sequences (sequences that share a common ancestry) in the aligned genomes and checking to what extent those sequences are conserved. Based on these, genome and molecular evolution are inferred and this may in turn be put in the context of, for example, phenotypic evolution or population genetics.

Virtually started as soon as the whole genomes of two organisms became available (that is, the genomes of the bacteria *Haemophilus influenzae* and *Mycoplasma genitalium*) in 1995, comparative genomics is now a standard component of the analysis of every new genome sequence. With the explosion in the number of genome project due to the advancements in DNA sequencing technologies, particularly the next-generation sequencing methods in late 2000s, this field has become more sophisticated, making it possible to deal with many genomes in a single study. Comparative genomics has revealed high levels of similarity between closely related organisms, such as humans and chimpanzees, and, more surprisingly, similarity between seemingly distantly related organisms, such as humans and the yeast *Saccharomyces cerevisiae*. It has also showed the extreme diversity of the gene composition in different evolutionary lineages

Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved (selection is neutral).

Identifying the mechanisms of eukaryotic genome evolution by comparative genomics is one of the important goals of the field. It is however often complicated by the multiplicity of events that have taken place throughout the history of individual lineages, leaving only distorted and superimposed traces in the genome of each living organism. For

this reason comparative genomics studies of small model organisms (for example yeast) are of great importance to advance our understanding of general mechanisms of evolution.

Having come a long way from its initial use of finding functional proteins, comparative genomics is now concentrating on finding regulatory regions and siRNA molecules. Recently, it has been discovered that distantly related species often share long conserved stretches of DNA that do not appear to code for any protein. It is unknown at this time what function such ultra-conserved regions serve.

Evolutionary principles

One character of biology is evolution, evolutionary theory is also the theoretical foundation of comparative genomics, and at the same time the results of comparative genomics unprecedentedly enriched and developed the theory of evolution. When two or more of the genome sequence compared, in essence get the evolutionary relationships of the sequence in the phylogenetic tree. Increased genome information of study makes molecular evolution, gene function at the genome level possible. Based on a variety of biological genome data and the study of vertical and horizontal evolution process, can understand vital parts of gene structure and its regulation function for life. But as a result in biological genome about 1.5% ~ 14.5% of the genes related to "lateral migration phenomenon", namely the gene transfer between populations which can exist at the same time, the differences in sequence result has nothing to do with the evolution. So in the system analysis, it needs to establish a relatively complete evolution model, in order to avoid gene transfer and the influence of the lack of more appropriate species which are conserved sequence.

Similarity of related genomes is the basis of comparative genomics. Two creatures which have a recent common ancestor, the species difference genomes between them were evolved from ancestors' genome, the closer the two organisms on the evolutionary stages, the higher their genome correlated. If there is close relationship between them, then their genome will behave like linear (synteny), namely some or all of the genetic sequences are conservative. So scientists can use the homology of the sequence and structure of encoding between mode genomes, by known genome mapping information to locate other genes in the genome, so as to reveal the potential function of the genes, clarify evolutionary relationship and the inner structure of the genome.

Orthologous sequences are separate because of speciation: a gene exists in the original species, the species divided into two species, so genes in new species are orthologous. Paralogy sequences are separate by gene cloning (gene duplication): if a particular gene in the biology is copied, then the copy of the two sequences is paralogy. A pair of orthologous

sequences is called orthologous pairs (orthologs), a pair of paralogy sequence is called collateral pairs (paralogs). Orthologous pairs usually have the same or similar function, but not necessarily on collateral pairs: due to the lack of the power of natural selection, the original duplicate copy of the genes are variation and get free new functions.

Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved (selection is neutral).

One of the important goals of the field is the identification of the mechanisms of eukaryotic genome evolution. It is however often complicated by the multiplicity of events that have taken place throughout the history of individual lineages, leaving only distorted and superimposed traces in the genome of each living organism. For this reason comparative genomics studies of small model organisms (for example the model *Caenorhabditis elegans* and closely related *Caenorhabditis briggsae*) are of great importance to advance our understanding of general mechanisms of evolution

Tools

Computational tools for analyzing sequences and complete genomes are developed quickly due to the availability of large amount of genomic data. At the same time, comparative analysis tools are progressed and improved. In the challenges about these analyses, it is very important to visualize the comparative results.

Visualization of sequence conservation is a tough task of comparative sequence analysis. As we know, it is highly inefficient to examine the alignment of long genomic regions manually. Internet-based genome browsers provide many useful tools for investigating genomic sequences due to integrating all sequence-based biological information on genomic regions. When we extract large amount of relevant biological data, they can be very easy to use and less time-consuming.

UCSC Browser: This site contains the reference sequence and working draft assemblies for a large collection of genomes.[26]

Ensembl: The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

MapView: The Map Viewer provides a wide variety of genome mapping and sequencing data.

VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. It was built to visualize the results of comparative analysis based on DNA alignments. The presentation of comparative data generated by VISTA can easily suit both small and large scale of data.

An advantage of using online tools is that these websites are being developed and updated constantly. There are many new settings and content can be used online to improve efficiency.[25]

Applications

Gene identification

Once genome correspondence is established, comparative genomics can aid gene identification. Comparative genomics can recognize real genes based on their patterns of nucleotide conservation across evolutionary time. With the availability of genome-wide alignments across the genomes compared, the different ways by which sequences change in known genes and in intergenic regions can be analyzed. The alignments of known genes will reveal the conservation of the reading frame of protein translation.

The genome of a species encodes genes and other functional elements, interspersed with non-functional nucleotides in a single uninterrupted string of DNA. Recognizing protein-coding genes typically relies on finding stretches of nucleotides free of stop codons (called Open Reading Frames, or ORFs) that are too long to have likely occurred by chance. Since stop codons occur at a frequency of roughly 1 in 20 in random sequence, ORFs of at least 60 amino acids will occur frequently by chance (5% under a simple Poisson model), and even ORFs of 150 amino acids will appear by chance in a large genome (0.05%). This poses a huge challenge for higher eukaryotes in which genes are typically broken into many, small exons (on average 125 nucleotides long for internal exons) in mammals. The basic problem is distinguishing *real genes* – those ORFs encoding a translated protein product – from *spurious ORFs* – the remaining ORFs whose presence is simply due to chance. In mammalian genomes, estimates of hypothetical genes have ranged from 28,000 to more than 120,000 genes. The internal coding exons were easily identified using Comparative analysis of human genome with mouse genome.

Regulatory motif discovery

Regulatory motifs are short DNA sequences about 6 to 15bp long that are used to control the expression of genes, dictating the conditions under which a gene will be turned on

or off. Each motif is typically recognized by a specific DNA-binding protein called a transcription factor (TF). A transcription factor binds precise sites in the promoter region of target genes in a sequence-specific way, but this contact can tolerate some degree of sequence variation. Thus, different binding sites may contain slight variations of the same underlying motif, and the definition of a regulatory motif should capture these variations while remaining as specific as possible. Comparative genomics provides a powerful way to distinguish regulatory motifs from non-functional patterns based on their conservation. One such example is the identification of TF DNA-binding motif using comparative genomics and *denovo* motif. The regulatory motifs of the Human Promoters were identified by comparison with other mammals. Yet another important finding is the gene and regulatory element by comparison of yeast species.

Agriculture

Agriculture is a field that reaps the benefits of comparative genomics. Identifying the loci of advantageous genes is a key step in breeding crops that are optimized for greater yield, cost-efficiency, quality, and disease resistance. For example, one genome wide association study conducted on 517 rice landraces revealed 80 loci associated with several categories of agronomic performance, such as grain weight, amylose content, and drought tolerance. Many of the loci were previously uncharacterized. Not only is this methodology powerful, it is also quick. Previous methods of identifying loci associated with agronomic performance required several generations of carefully monitored breeding of parent strains, a time consuming effort that is unnecessary for comparative genomic studies.

Medicine

The medical field also benefits from the study of comparative genomics. Vaccinology in particular has experienced useful advances in technology due to genomic approaches to problems. In an approach known as reverse vaccinology, researchers can discover candidate antigens for vaccine development by analyzing the genome of a pathogen or a family of pathogens.[32] Applying a comparative genomics approach by analyzing the genomes of several related pathogens can lead to the development of vaccines that are multiprotective. A team of researchers employed such an approach to create a universal vaccine for Group B *Streptococcus*, a group of bacteria responsible for severe neonatal infection. Comparative genomics can also be used to generate specificity for vaccines against pathogens that are closely related to commensal microorganisms. For example, researchers used comparative genomic analysis of commensal and pathogenic strains of *E. coli* to identify pathogen specific

genes as a basis for finding antigens that result in immune response against pathogenic strains but not commensal ones.

Research

Comparative genomics also opens up new avenues in other areas of research. As DNA sequencing technology has become more accessible, the number of sequenced genomes has grown. With the increasing reservoir of available genomic data, the potency of comparative genomic inference has grown as well. A notable case of this increased potency is found in recent primate research. Comparative genomic methods have allowed researchers to gather information about genetic variation, differential gene expression, and evolutionary dynamics in primates that were indiscernible using previous data and methods. The Great Ape Genome Project used comparative genomic methods to investigate genetic variation with reference to the six great ape species, finding healthy levels of variation in their gene pool despite shrinking population size. Another study showed that patterns of DNA methylation, which are a known regulation mechanism for gene expression, differ in the prefrontal cortex of humans versus chimps, and implicated this difference in the evolutionary divergence of the two species

Other applications

Comparative genomics has wide applications in the field of molecular medicine and molecular evolution. The most significant application of comparative genomics in molecular medicine is the identification of drug targets of many infectious diseases. For example, comparative analyses of fungal genomes have led to the identification of many putative targets for novel antifungal. This discovery can aid in target based drug design to cure fungal diseases in human. Comparative analysis of genomes of individuals with genetic disease against healthy individuals may reveal clues of eliminating that disease.

Comparative genomics helps in selecting model organisms. A model system is a simple, idealized system that can be accessible and easily manipulated. For example, a comparison of the fruit fly genome with the human genome discovered that about 60 percent of genes are conserved between fly and human. Researchers have found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. Even more surprisingly, when scientists inserted a human gene associated with early-onset Parkinson's disease into fruit flies, they displayed symptoms similar to those seen in humans with the disorder, raising the possibility that the tiny insects could serve as a new model for testing therapies aimed at Parkinson's. Thus, comparative genomics may provide gene functional

annotation. Gene finding is an important application of comparative genomics. Comparative genomics identify Synteny (genes present in the same order in the genomes) and hence reveal gene clusters.

Comparative genomics also helps in the clustering of regulatory sites [10], which can help in the recognition of unknown regulatory regions in other genomes. The metabolic pathway regulation can also be recognized by means of comparative genomics of a species. Dmitry and colleagues have identified the regulons of methionine metabolism in gram-positive bacteria using comparative genomics analysis. Similarly Kai Tan [12] and colleagues have identified regulatory networks of *H. influenzae* by comparing its genome with that of *E. coli*. The adaptive properties of organisms [13] like evolution of sex, gene silencing can also be correlated to genome sequence by comparative genomics.

Metabolomics

Metabolomics is the scientific study of chemical processes involving metabolites. Specifically, metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles.[1] The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes. mRNA gene expression data and proteomic analyses reveal the set of gene products being produced in the cell, data that represents one aspect of cellular function. Conversely, metabolic profiling can give an instantaneous snapshot of the physiology of that cell. One of the challenges of systems biology and functional genomics is to integrate proteomic, transcriptomic, and metabolomic information to provide a better understanding of cellular biology.

History

The idea that biological fluids reflect the health of an individual has existed for a long time. Ancient Chinese doctors used ants for the evaluation of urine of patients to detect whether the urine contained high levels of glucose, and hence detect diabetes.[3] In the Middle Ages, "urine charts" were used to link the colours, tastes and smells of urine to various medical conditions, which are metabolic in origin.[4]

The concept that individuals might have a "metabolic profile" that could be reflected in the makeup of their biological fluids was introduced by Roger Williams in the late 1940s,[5] who used paper chromatography to suggest characteristic metabolic patterns in urine and saliva were associated with diseases such as schizophrenia. However, it was only

through technological advancements in the 1960s and 1970s that it became feasible to quantitatively (as opposed to qualitatively) measure metabolic profiles.[6] The term "metabolic profile" was introduced by Horning, et al. in 1971 after they demonstrated that gas chromatography-mass spectrometry(GC-MS) could be used to measure compounds present in human urine and tissue extracts.[3][7] The Horning group, along with that of Linus Pauling and Arthur B. Robinson led the development of GC-MS methods to monitor the metabolites present in urine through the 1970s.[8]

Concurrently, NMR spectroscopy, which was discovered in the 1940s, was also undergoing rapid advances. In 1974, Seeley et al. demonstrated the utility of using NMR to detect metabolites in unmodified biological samples.[9] This first study on muscle highlighted the value of NMR in that it was determined that 90% of cellular ATP is complexed with magnesium. As sensitivity has improved with the evolution of higher magnetic field strengths and magic angle spinning, NMR continues to be a leading analytical tool to investigate metabolism.[3][4] Recent[when?] efforts to utilize NMR for metabolomics have been largely driven by the laboratory of Jeremy K. Nicholson at Birkbeck College, University of London and later at Imperial College London. In 1984, Nicholson showed ¹H NMR spectroscopy could potentially be used to diagnose diabetes mellitus, and later pioneered the application of pattern recognition methods to NMR spectroscopic data.[10][11]

In 2005, the first metabolomics web database, METLIN,[12] for characterizing human metabolites was developed in the Siuzdak laboratory at The Scripps Research Institute and contained over 10,000 metabolites and tandem mass spectral data. As of September 2015, METLIN contains over 240,000 metabolites as well as the largest repository of tandem mass spectrometry data in metabolomics.

On 23 January 2007, the Human Metabolome Project, led by Dr. David Wishart of the University of Alberta, Canada, completed the first draft of the human metabolome, consisting of a database of approximately 2500 metabolites, 1200 drugs and 3500 food components.[13][14] Similar projects have been underway in several plant species, most notably *Medicago truncatula*[15] and *Arabidopsis thaliana*[16] for several years.

As late as mid-2010, metabolomics was still considered an "emerging field".[17] Further, it was noted that further progress in the field depended in large part, through addressing otherwise "irresolvable technical challenges", by technical evolution of mass spectrometry instrumentation.[17]

In 2015, real-time metabolome profiling was demonstrated for the first time.[18]

Metabolome

Metabolome refers to the complete set of small-molecule metabolites (such as metabolic intermediates, hormones and other signaling molecules, and secondary metabolites) to be found within a biological sample, such as a single organism.[19][20] The word was coined in analogy with transcriptomics and proteomics; like the transcriptome and the proteome, the metabolome is dynamic, changing from second to second. Although the metabolome can be defined readily enough, it is not currently possible to analyse the entire range of metabolites by a single analytical method. The first metabolite database (called METLIN) for searching m/z values from mass spectrometry data was developed by scientists at The Scripps Research Institute in 2005.[12] In January 2007, scientists at the University of Alberta and the University of Calgary completed the first draft of the human metabolome. They catalogued approximately 2500 metabolites, 1200 drugs and 3500 food components that can be found in the human body, as reported in the literature.[13] This information, available at the Human Metabolome Database (www.hmdb.ca) and based on analysis of information available in the current scientific literature, is far from complete.[21] In contrast, much more is known about the metabolomes of other organisms. For example, over 50,000 metabolites have been characterized from the plant kingdom, and many thousands of metabolites have been identified and/or characterized from single plants.

Each type of cell and tissue has a unique metabolic 'fingerprint' that can elucidate organ or tissue-specific information, while the study of biofluids can give more generalized though less specialized information. Commonly used biofluids are urine and plasma, as they can be obtained non-invasively or relatively non-invasively, respectively.[24] The ease of collection facilitates high temporal resolution, and because they are always at dynamic equilibrium with the body, they can describe the host as a whole.[25]

Metabolites

Metabolites are the intermediates and products of metabolism. Within the context of metabolomics, a metabolite is usually defined as any molecule less than 1 kDa in size.[26] However, there are exceptions to this depending on the sample and detection method. For example, macromolecules such as lipoproteins and albumin are reliably detected in NMR-based metabolomics studies of blood plasma.[27] In plant-based metabolomics, it is common to refer to "primary" and "secondary" metabolites. A primary metabolite is directly involved in the normal growth, development, and reproduction. A secondary metabolite is not directly involved in those processes, but usually has important ecological function. Examples include antibiotics and pigments.[28] By contrast, in human-based metabolomics, it is more

common to describe metabolites as being either endogenous (produced by the host organism) or exogenous.[29] Metabolites of foreign substances such as drugs are termed xenometabolites.[30]

The metabolome forms a large network of metabolic reactions, where outputs from one enzymatic chemical reaction are inputs to other chemical reactions. Such systems have been described as hypercycles.

Metabonomics

Metabonomics is defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification". The word origin is from the Greek *μεταβολή* meaning change and *nomos* meaning a rule set or set of laws.[31] This approach was pioneered by Jeremy Nicholson at Imperial College London and has been used in toxicology, disease diagnosis and a number of other fields. Historically, the metabonomics approach was one of the first methods to apply the scope of systems biology to studies of metabolism. There has been some disagreement over the exact differences between 'metabolomics' and 'metabonomics'. The difference between the two terms is not related to choice of analytical platform: although metabonomics is more associated with NMR spectroscopy and metabolomics with mass spectrometry-based techniques, this is simply because of usages amongst different groups that have popularized the different terms. While there is still no absolute agreement, there is a growing consensus that 'metabolomics' places a greater emphasis on metabolic profiling at a cellular or organ level and is primarily concerned with normal endogenous metabolism. 'Metabonomics' extends metabolic profiling to include information about perturbations of metabolism caused by environmental factors (including diet and toxins), disease processes, and the involvement of extragenomic influences, such as gut microflora. This is not a trivial difference; metabolomic studies should, by definition, exclude metabolic contributions from extragenomic sources, because these are external to the system being studied. However, in practice, within the field of human disease research there is still a large degree of overlap in the way both terms are used, and they are often in effect synonymous.[35]

Designing a metabolomics study

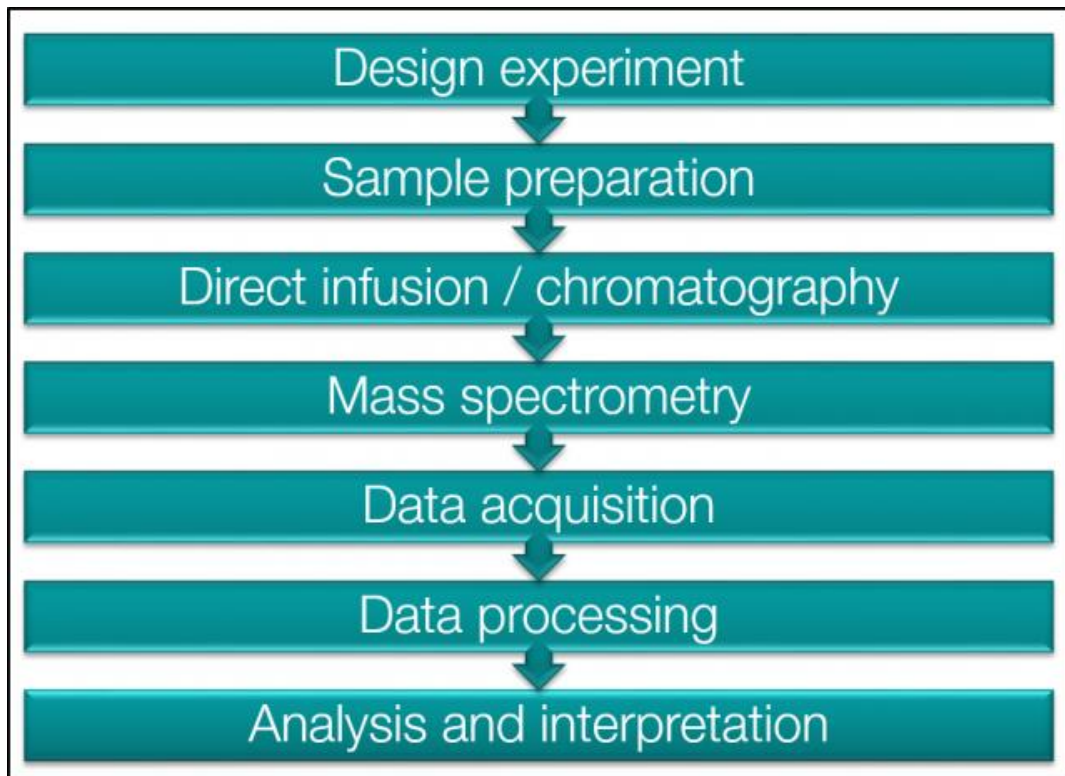
The two main approaches that can be used in *metabolomics* are untargeted and targeted approaches.

The approach chosen will determine how you design your experiment, prepare your samples, and what analytical techniques you use.

i **Untargeted (global) approach:** This method measures as many metabolites as possible from a range of biological samples without any (intended) bias.

i **Targeted approach:** This method is used when you want to measure sets of metabolites and have a specific biochemical question that you want to answer.

This approach is often used in pharmacokinetic studies of drug *metabolism* and when looking at the effect of therapeutics or genetic modifications on a specific enzyme.



Analytical technologies

Separation methods

Initially, analytes in a metabolomic sample comprise highly complex mixture. This complex mixture can be simplified prior to detection by separating some analytes from others. Separation achieves various goals: analytes which cannot be resolved by the detector may be separated in this step; in MS analysis ion suppression is reduced; the retention time of the analyte serves as information regarding its identity. This separation step is not mandatory and is often omitted in NMR and "shotgun" based approaches such as shotgun lipidomics.

- Gas chromatography, especially when interfaced with mass spectrometry (GC-MS), is one of the most widely used methods for metabolomic analysis.[citation needed] GC offers very high chromatographic resolution, but requires chemical derivatization for many biomolecules as only volatile

chemicals can be analysed without derivatization. In cases where greater separation is required, Comprehensive Chromatography (GCxGC) also can be applied.

- High performance liquid chromatography (HPLC) is another very common method for metabolomic analysis. With the advent of electrospray ionization, HPLC was coupled to MS. As compared to GC, HPLC has lower chromatographic resolution, but requires no derivitization for polar molecules and separates molecules in the liquid phase. Additionally HPLC has the advantage that a much wider range of analytes can be measured with a higher sensitivity than GC methods.[36]
- Capillary electrophoresis (CE). CE has a higher theoretical separation efficiency than HPLC, and is suitable for use with a wider range of metabolite classes than is GC. As for all electrophoretic techniques, it is most appropriate for charged analytes.[37]

Ionization methods

For analysis by mass spectrometry the analytes must be imparted with a charge and transferred to the gas phase.

- Electron ionization (EI) is the most common ionization technique applies to GC separations as it is amenable to low pressures. EI also produces fragmentation of the analyte, both providing structural information while increasing the complexity of the data and possibly obscuring the molecular ion.
- Chemical ionization (CI) is an atmospheric pressure technique that can be applied to all the above separation techniques. CI maintains the molecular ion while providing a more aggressive ionization than ESI which is suitable for less polar compounds.
- Electrospray ionization (ESI) is the most common ionization technique applied in LC/MS. This soft ionization is most successful for polar molecules with ionizable functional groups.

Detection methods

Mass spectrometry

Mass spectrometry (MS) is used to identify and to quantify metabolites after optional separation by GC, HPLC (LC-MS), or CE. GC-MS was the first hyphenated technique to be

developed. Identification leverages the distinct patterns in which analytes fragment which can be thought of as a mass spectral fingerprint; libraries exist that allow identification of a metabolite according to this fragmentation pattern. MS is both sensitive and can be very specific. There are also a number of techniques which use MS as a stand-alone technology: the sample is infused directly into the mass spectrometer with no prior separation, and the MS provides sufficient selectivity to both separate and to detect metabolites.

Surface-based mass analysis

Surface-based mass analysis has seen a resurgence in the past decade, with new MS technologies focused on increasing sensitivity, minimizing background, and reducing sample preparation. The ability to analyze metabolites directly from biofluids and tissues continues to challenge current MS technology, largely because of the limits imposed by the complexity of these samples, which contain thousands to tens of thousands of metabolites. Among the technologies being developed to address this challenge is Nanostructure-Initiator MS (NIMS),[38][39] a desorption/ ionization approach that does not require the application of matrix and thereby facilitates small-molecule (i.e., metabolite) identification. MALDI is also used however, the application of a MALDI matrix can add significant background at <1000 Da that complicates analysis of the low-mass range (i.e., metabolites). In addition, the size of the resulting matrix crystals limits the spatial resolution that can be achieved in tissue imaging. Because of these limitations, several other matrix-free desorption/ionization approaches have been applied to the analysis of biofluids and tissues. Secondary ion mass spectrometry (SIMS) was one of the first matrix-free desorption/ionization approaches used to analyze metabolites from biological samples. SIMS uses a high-energy primary ion beam to desorb and generate secondary ions from a surface. The primary advantage of SIMS is its high spatial resolution (as small as 50 nm), a powerful characteristic for tissue imaging with MS. However, SIMS has yet to be readily applied to the analysis of biofluids and tissues because of its limited sensitivity at >500 Da and analyte fragmentation generated by the high-energy primary ion beam. Desorption electrospray ionization (DESI) is a matrix-free technique for analyzing biological samples that uses a charged solvent spray to desorb ions from a surface. Advantages of DESI are that no special surface is required and the analysis is performed at ambient pressure with full access to the sample during acquisition. A limitation of DESI is spatial resolution because "focusing" the charged solvent spray is difficult. However, a recent development termed laser ablation ESI (LAESI) is a promising approach to circumvent this limitation.

Nuclear magnetic resonance (NMR) spectroscopy.

NMR is the only detection technique which does not rely on separation of the analytes, and the sample can thus be recovered for further analyses. All kinds of small molecule metabolites can be measured simultaneously - in this sense, NMR is close to being a universal detector. The main advantages of NMR are high analytical reproducibility and simplicity of sample preparation. Practically, however, it is relatively insensitive compared to mass spectrometry-based techniques.[40][41] Another study demonstrated that the lack of sensitivity of the current NMR based metabolomics protocol is not due to instrumental limitations but rather due to methodological limitations. In the same way that binning of spectra negates the better resolution of higher field magnets; the same appears to be true by employing PCA analysis to NMR spectra which effectively negates the higher sensitivity of higher field magnets.[42]

Although NMR and MS are the most widely used, modern day techniques other methods of detection that have been used. These include ion-mobility spectrometry, electrochemical detection (coupled to HPLC) and radiolabel (when combined with thin-layer chromatography)

Statistical methods

The data generated in metabolomics usually consist of measurements performed on subjects under various conditions. These measurements may be digitized spectra, or a list of metabolite levels. In its simplest form this generates a matrix with rows corresponding to subjects and columns corresponding with metabolite levels.[3] Several statistical programs are currently available for analysis of both NMR and mass spectrometry data. For mass spectrometry data, software is available that identifies molecules that vary in subject groups on the basis of mass and sometimes retention time depending on the experimental design. The first comprehensive software to analyze global mass spectrometry-based metabolomics datasets was developed by the Siuzdak laboratory at The Scripps Research Institute in 2006. This software, called XCMS, is freely available, has over 20,000 downloads since its inception in 2006,[43] and is one of the most widely cited mass spectrometry-based metabolomics software programs in scientific literature. XCMS has now been surpassed in usage by a cloud-based version of XCMS called XCMS Online.[44][45] Other popular metabolomics programs for mass spectral analysis are MZmine,[46] MetAlign,[47] MathDAMP,[48] which also compensate for retention time deviation during sample analysis. LCMStats[49] is another R package for detailed analysis of liquid chromatography mass spectrometry(LCMS)data and is helpful in identification of co-eluting ions especially isotopologues from a complicated metabolic profile. It combines xcms

package functions and can be used to apply many statistical functions for correcting detector saturation using coates correction and creating heat plots. Metabolomics data may also be analyzed by statistical projection (chemometrics) methods such as principal components analysis and partial least squares regression.[50]

Once metabolic composition is determined, data reduction techniques can be used to elucidate patterns and connections. In many studies, including those evaluating drug-toxicity and some disease models, the metabolites of interest are not known a priori. This makes unsupervised methods, those with no prior assumptions of class membership, a popular first choice. The most common of these methods includes principal component analysis (PCA) which can efficiently reduce the dimensions of a dataset to a few which explain the greatest variation[25] When analyzed in the lower-dimensional PCA space, clustering of samples with similar metabolic fingerprints can be detected. This clustering can elucidate patterns and assist in the determination of disease biomarkers - metabolites that correlate most with class membership.

Key applications

Toxicity assessment/toxicology.

Metabolic profiling (especially of urine or blood plasma samples) detects the physiological changes caused by toxic insult of a chemical (or mixture of chemicals). In many cases, the observed changes can be related to specific syndromes, e.g. a specific lesion in liver or kidney. This is of particular relevance to pharmaceutical companies wanting to test the toxicity of potential drug candidates: if a compound can be eliminated before it reaches clinical trials on the grounds of adverse toxicity, it saves the enormous expense of the trials.

Functional genomics.

Metabolomics can be an excellent tool for determining the phenotype caused by a genetic manipulation, such as gene deletion or insertion. Sometimes this can be a sufficient goal in itself—for instance, to detect any phenotypic changes in a genetically modified plant intended for human or animal consumption. More exciting is the prospect of predicting the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes. Such advances are most likely to come from model organisms such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. The Cravatt laboratory at The Scripps Research Institute has recently applied this technology to mammalian systems, identifying the N-acyltaurines as previously uncharacterized endogenous substrates for the enzyme fatty acid amide hydrolase (FAAH) and the

Omics concepts

monoalkylglycerol ethers (MAGEs) as endogenous substrates for the uncharacterized hydrolase KIAA1363.

Nutrigenomics

Nutrigenomics is a generalised term which links genomics, transcriptomics, proteomics and metabolomics to human nutrition. In general a metabolome in a given body fluid is influenced by endogenous factors such as age, sex, body composition and genetics as well as underlying pathologies. The large bowel microflora are also a very significant potential confounder of metabolic profiles and could be classified as either an endogenous or exogenous factor. The main exogenous factors are diet and drugs. Diet can then be broken down to nutrients and non-nutrients. Metabolomics is one means to determine a biological endpoint, or metabolic fingerprint, which reflects the balance of all these forces on an individual's metabolism.[53]

Agricultural

The development of new pesticides is critical to meet the growing demands on farming. Metabolomics enables us to improve genetically modified plants, and helps us to estimate associated risks by allowing us to get a glimpse of their complex biochemistry via informative snapshots acquired at different time points during plant development.

Plant *metabolomics* is particularly interesting because of the range and functions of primary and secondary metabolites in plants. About 300 distinct metabolites could be routinely identified per sample a decade ago, and the number is gradually increasing over time.

Biomarker discovery

Biomarker discovery is another area where metabolomics informs decision making. Biomarkers are "objective indications of medical state observed from outside the patient - which can be measured accurately and reproducibly" (Kyle *et al.* What are Biomarkers?). In metabolomics, biomarkers are *small molecules* (metabolites) that can be used to distinguish two groups of samples, typically a disease and control group. For example, a metabolite reliably present in disease samples, but not in healthy individuals would be classed as a biomarker. Samples of urine, saliva, bile, or seminal fluid contain highly informative metabolites, and can be readily analysed through metabolomics fingerprinting or profiling, for the purpose of biomarker discovery.

Personalised medicine

Personalised medicine, the ultimate customisation of healthcare, requires metabolomics for quick medical diagnosis to identify disease. In healthcare, we currently use

classical biochemical tests to measure individual metabolite concentrations to identify disease states (e.g. the blood-glucose level in the case of diabetes). Metabolomics offers the potential for the rapid identification of hundreds of metabolites, enabling us to identify these disease states much earlier.

Environmental metabolomics

Environmental metabolomics is the application of metabolomics to characterise the interactions of organisms with their environment. This approach has many advantages for studying organism–environment interactions and for assessing organism function and health at the molecular level. As such, metabolomics is finding an increasing number of applications in the environmental sciences, ranging from understanding organismal responses to abiotic pressures, to investigating the responses of organisms to other biota. These interactions can be studied from individuals to populations, which can be related to the traditional fields of ecophysiology and ecology, and from instantaneous effects to those over evolutionary time scales, the latter enabling studies of genetic adaptation.

Population Genomics

Population genomics — an emerging discipline and a new paradigm in population genetics¹ — combines genomic concepts and technologies with the population genetics objective of understanding evolution. The term was apparently first used in a publication about human disease genetics by Gulcher and Stefansson² and subsequently has become increasingly popular. Population genomics can be broadly defined as the simultaneous study of numerous loci or genome regions to better understand the roles of evolutionary processes (such as mutation, RANDOM GENETIC DRIFT, GENE FLOW and natural selection) that influence variation across genomes and populations. This broad definition includes issues ranging from understanding the pattern and degree of genome-wide heterogeneity (for example, chromosomal/positional differences in sequence diversity and recombination rates) to the origins, relationships and demographic history (interpopulation movement rates, relationships and relative divergence dates) of populations using genome-wide sampling.

According to a more narrow definition, as proposed by Black et al.¹, population genomics is the use of genome-wide sampling to identify and to separate locus specific effects (such as selection, mutation, assortative mating and recombination) from genome-wide effects (such as drift or BOTTLENECKS, gene flow and inbreeding) to improve our understanding of MICROEVOLUTION. This is crucial because only genome-wide effects inform us reliably about population demography and phylogenetic history, whereas locus-specific effects help identify genes that are important for fitness and adaptation.

An example of a locus-specific effect is directional selection whereby one allele is selected for in population X but another is selected for in population Y. Such selection would generate a large allele frequency difference (high F_{ST}) at this locus relative to the F_{ST} at distant non-linked NEUTRAL LOCI across the genome. The two main principles of population genomics are that neutral loci across the genome will be similarly affected by demography and the evolutionary history of populations, and that loci under selection will often behave differently and therefore reveal ‘outlier’ patterns of variation. Consequently, it is extremely important to identify OUTLIER LOCI both to reliably infer population demographic history (in which case outliers often should be excluded) and to detect selected (adaptive) loci. Selection will also influence linked markers along a chromosome, such that a SELECTION SIGNATURE (outlier effects) can often be detected by genotyping markers that are scattered across chromosomes (even if the marker is not in the gene that is affected by selection). The selection signature will decay with time owing to recombination, and therefore ancient/historical selection might not be detectable.

Here, we outline four steps that constitute a basic population-genomic approach, which is based on genotyping numerous molecular markers and testing for outlier loci in population data sets. We illustrate the concept of ‘outlier loci’, discuss recent statistical and molecular genomic approaches that detect outliers (including available computer software), and quantify the magnitude of bias caused by outliers when estimating POPULATION PARAMETERS (for example, migration rates). Complementary population-genomic approaches, such as quantitative trait loci (QTL) mapping in controlled populations, population-based mapping of genes through linkage disequilibrium (LD) and association studies, have been reviewed elsewhere^{8–10} and are not discussed here. We conclude with a brief discussion of some important uses of outlier markers for biodiversity conservation; other uses (such as detecting SELECTIVE SWEEPS) have been reviewed elsewhere^{11–14}. We focus mainly on non-human and non-model organisms because the increasing availability of genome-scale data sets in this more diverse set of taxa will yield evolutionary insights that are more broadly applicable. In addition, further genome-wide thinking is needed in studies of non-model taxa to improve evolutionary inferences. Readers who are interested in human and medical genomics (pharmacogenomics) should consult. We also concentrate on molecular marker data — microsatellites, single nucleotide polymorphisms (SNPs) and AMPLIFIED FRAGMENT-LENGTH POLYMORPHISMS (AFLPs) — as they will continue to be the most widely useful markers for non-model organisms in the near future.

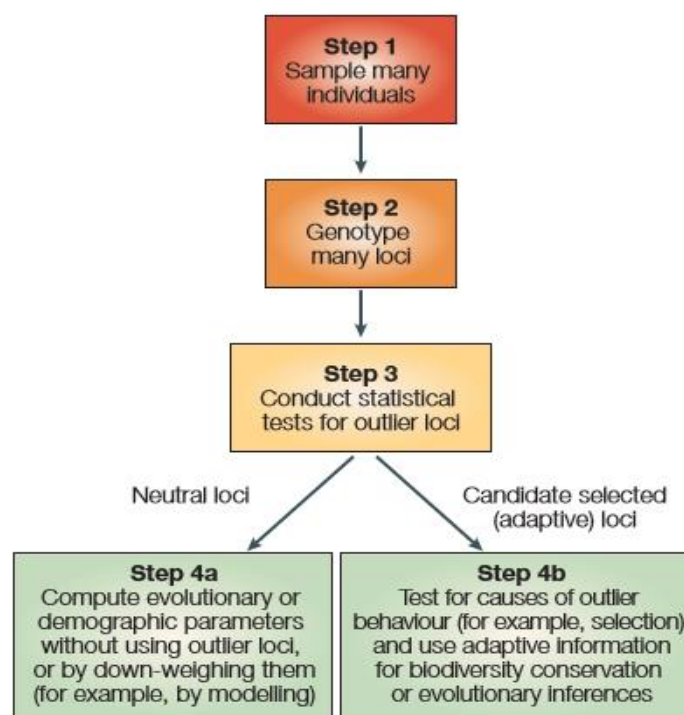


Figure 1 |
the four main
population-
approach. The
summarized here

Flow chart of
steps in the
genomic
approach
can be used to

identify loci that are under selection (adaptive genes) and to better estimate population

history and demography. Step 1, if searching for adaptive variation, is to sample groups of individuals with divergent phenotypes or to sample across a ‘selection gradient’ (for example, in disease exposure, environmental conditions or phenotype). Large populations are sampled because selection signatures will be detectable only if they are not obscured by drift (small effective population size, N_e). The selection coefficient (s) must be large relative to the N_e for selection to be detectable (for example, $(N_e \times s) > 1$; see REF.33). Step 2 is to conduct genome-wide genotyping, preferably with mapped loci. If inferring demographic status, independent neutral loci (for example, pseudogenes, random markers and non-coding sequences) are used. If searching for adaptive variation, markers that are in or near genes (ideally of known function and associated with phenotype or environment) are used, as well as many neutral markers. In step 3, outlier loci are those that behave unlike most other loci in the sample; for example, those with an extremely high F_{st} (genetic divergence). Such loci are potentially under selection and could mark adaptive variation; they could also bias estimates of parameters such as gene flow, population size and structure, and therefore should not be used (or be accounted for by modelling). The key to improving many applications of molecular markers in population genetics is the development and validation of improved statistical tests to identify and deal with outlier loci. Step 4a is to estimate N_e , N_m (dispersal rate), F_{st} , structure and phylogenies, to test for bottlenecks, expansion, Hardy–Weinberg and genotypic disequilibrium, and so on. Step 4b is to validate selection as the cause of outlier behaviour — for example, by correlating patterns at outlier loci with selection gradients of environmental variables⁶ (disease presence, temperature gradients, predation and so on). Selected markers should be used in studies to better understand adaptation or to plan conservation-management strategies.

Fluxomics

Fluxomics describes the various approaches that seek to determine the rates of metabolic reactions within a biological entity. While, metabolomics can provide instantaneous information on the metabolites in a biological sample, metabolism is a dynamic process. The significance of fluxomics is that metabolic fluxes determine the cellular phenotype. It has the added advantage of being based on the metabolome which has fewer components than the genome or proteome.

Fluxomics falls within the field of systems biology which developed with the appearance of high throughput technologies. Systems biology recognizes the complexity of biological systems and has the broader goal of explaining and predicting this complex behavior.

Metabolic flux

Metabolic flux refers to the rate of metabolite conversion in a metabolic network. For a reaction this rate is a function of both enzyme abundance and enzyme activity. Enzyme concentration is itself a function of transcriptional and translational regulation in addition to the stability of the protein. Enzyme activity is affected by the kinetic parameters of the enzyme, the substrate concentrations, the product concentrations, and the effector molecules concentration. The genomic and environmental effects on metabolic flux are what determine healthy or diseased phenotype.

Fluxome

Similar to genome, transcriptome, proteome, and metabolome, the fluxome is defined as the complete set of metabolic fluxes in a cell. However, unlike the others the fluxome is a dynamic representation of the phenotype. This is due to the fluxome resulting from the interactions of the metabolome, genome, transcriptome, proteome, post-translational modifications and the environment.

Flux analysis technologies

Two important technologies are flux balance analysis (FBA) and ^{13}C -fluxomics. In FBA metabolic fluxes are estimated by first representing the metabolic reactions of a metabolic network in a numerical matrix containing the stoichiometric coefficients of each reaction. The stoichiometric coefficients constrain the system model and are why FBA is only applicable to steady state conditions. Additional constraints can be imposed. By providing constraints the possible set of solutions to the system are reduced. Following the addition of constraints the system model is optimized. Flux-balance analysis resources include the BIGG database, the COBRA toolbox, and FASIMU.

In ^{13}C -fluxomics, metabolic precursors are enriched with ^{13}C before being introduced to the system. Using an imaging technique such as mass spectrometry or nuclear magnetic resonance spectroscopy the level of incorporation of ^{13}C into metabolites can be measured and with stoichiometry the metabolic fluxes can be estimated.

Stoichiometric and kinetic paradigms

A number of different methods, broadly divided into stoichiometric and kinetic paradigms.

Within the stoichiometric paradigm, a number of relatively simple linear algebra methods utilise restricted metabolic networks or genome-scale metabolic network models to perform flux balance analysis and the array of techniques derived from it. These linear equations are useful for steady state conditions. Dynamic methods are not yet usable. On the more experimental side, metabolic flux analysis allows the empirical estimation of reaction rates by stable isotope labelling.

Within the kinetic paradigm, kinetic modelling of metabolic networks can be purely theoretical, exploring the potential space of dynamic metabolic fluxes under perturbations away from steady state using formalisms such as biochemical systems theory. Such explorations are most informative when accompanied by empirical measurements of the system under study following actual perturbations, as is the case in metabolic control analysis.

Constraint based reconstruction and analysis

Collected methods in fluxomics have been described as "COBRA" methods, for COntstraint Based Reconstruction and Analysis. A number of software tools and environments have been created for this purpose. Although it can only be measured indirectly, metabolic flux is the critical link between genes, proteins and the observable phenotype. This is due to the fluxome integrating mass-energy, information, and signaling networks. Fluxomics has the potential to provide a quantifiable representation of the effect the environment has on the phenotype because the fluxome describes the genome environment interaction.^[21] In the fields of metabolic engineering and systems biology, fluxomic methods are considered a key enabling technology due to their unique position in the ontology of biological processes, allowing genome scale stoichiometric models to act as a framework for the integration of diverse biological datasets.

Examples of use in research

One potential application of fluxomic techniques is in drug design. Rama et al.^[25] used FBA to study the Mycolic Acid Pathway in *Mycobacterium tuberculosis*. Mycolic acids are known to be important to *M. tuberculosis* survival and as such its pathway has been studied extensively. This allowed the construction of a model of the pathway and for FBA to analyze it. The results of this found multiple possible drug targets for future investigation.

FBA was used to analyze the metabolic networks of multidrug-resistant *Staphylococcus aureus*. By performing in silico single and double gene deletions many enzymes essential to growth were identified.

Glycomics

Glycomics is the comprehensive study of glycomes (the entire complement of sugars, whether free or present in more complex molecules of an organism), including genetic, physiologic, pathologic, and other aspects. Glycomics "is the systematic study of all glycan structures of a given cell type or organism" and is a subset of glycobiology. The term glycomics is derived from the chemical prefix for sweetness or a sugar, "glyco-", and was formed to follow the naming convention established by genomics (which deals with genes) and proteomics (which deals with proteins).

Challenges

-
- The complexity of sugars: regarding their structures, they are not linear instead they are highly branched. Moreover, glycans can be modified (modified sugars), this increases its complexity.
 - Complex biosynthetic pathways for glycans.
 - Usually glycans are found either bound to protein (glycoprotein) or conjugated with lipids (glycolipids).
 - Unlike genomes, glycans are highly dynamic.

This area of research has to deal with an inherent level of complexity not seen in other areas of applied biology. 68 building blocks (molecules for DNA, RNA and proteins; categories for lipids; types of sugar linkages for saccharides) provide the structural basis for the molecular choreography that constitutes the entire life of a cell. DNA and RNA have four building blocks each (the nucleosides or nucleotides). Lipids are divided into eight categories based on ketoacyl and isoprene. Proteins have 20 (the amino acids). Saccharides have 32 types of sugar linkages. While these building blocks can be attached only linearly for proteins and genes, they can be arranged in a branched array for saccharides, further increasing the degree of complexity.

Add to this the complexity of the numerous proteins involved, not only as carriers of carbohydrate, the glycoproteins, but proteins specifically involved in binding and reacting with carbohydrate:

- Carbohydrate-specific enzymes for synthesis, modulation, and degradation
- Lectins, carbohydrate-binding proteins of all sorts
- Receptors, circulating or membrane-bound carbohydrate-binding receptors

Importance

To answer this question one should know the different and important functions of glycans.

The following are some of those functions:

- Glycoproteins found on the cell surface play a critical role in bacterial and viral recognition.
- They are involved in cellular signaling pathways and modulate cell function.
- They are important in innate immunity.
- They determine cancer development.
- They orchestrate the cellular fate, inhibit proliferation, regulate circulation and invasion.
- They affect the stability and folding of proteins.
- They affect the pathway and fate of glycoproteins.
- There are many glycan-specific diseases, often hereditary diseases.

There are important medical applications of aspects of glycomics:

- Lectins fractionate cells to avoid graft-versus-host disease in hematopoietic stem cell transplantation.
- Activation and expansion of cytolytic CD8 T cells in cancer treatment.

Glycomics is particularly important in microbiology because glycans play diverse roles in bacterial physiology. Research in bacterial glycomics could lead to the development of:

- novel drugs
- bioactive glycans
- glycoconjugate vaccines

TOOLS USED

The following are examples of the commonly used techniques in glycan analysis.

High-resolution mass spectrometry (MS) and high-performance liquid chromatography (HPLC)

The most commonly applied methods are MS and HPLC, in which the glycan part is cleaved either enzymatically or chemically from the target and subjected to analysis.^[6] In case of glycolipids, they can be analyzed directly without separation of the lipid component.

N-glycans from glycoproteins are analyzed routinely by high-performance-liquid-chromatography (reversed phase, normal phase and ion exchange HPLC) after tagging the reducing end of the sugars with a fluorescent compound (reductive labeling). A large variety of different labels were introduced in the recent years, where 2-aminobenzamide (AB),

anthranilic acid (AA), 2-aminopyridin (PA), 2-aminoacridone (AMAC) and 3-(acetylamino)-6-aminoacridine (AA-Ac) are just a few of them.

O-glycans are usually analysed without any tags, due to the chemical release conditions preventing them to be labeled.

Fractionated glycans from high-performance liquid chromatography (HPLC) instruments can be further analyzed by MALDI-TOF-MS(MS) to get further informations about structure and purity. Sometimes glycan pools are analyzed directly by mass spectrometry without prefractionation, although a discrimination between isobaric glycan structures is more challenging or even not always possible. Anyway, direct MALDI-TOF-MS analysis can lead to a fast and straightforward illustration of the glycan pool.

In recent years, high performance liquid chromatography online coupled to mass spectrometry became very popular. By choosing porous graphitic carbon as a stationary phase for liquid chromatography, even non derivatized glycans can be analyzed. Detection is here done by mass spectrometry, but instead of MALDI-MS, electrospray ionisation (ESI) is more frequently used.

Multiple Reaction Monitoring (MRM)

Although MRM has been used extensively in metabolomics and proteomics, its high sensitivity and linear response over a wide dynamic range make it especially suited for glycan biomarker research and discovery. MRM is performed on a triple quadrupole (QqQ) instrument, which is set to detect a predetermined precursor ion in the first quadrupole, a fragmented in the collision quadrupole, and a predetermined fragment ion in the third quadrupole. It is a non-scanning technique, wherein each transition is detected individually and the detection of multiple transitions occurs concurrently in duty cycles. This technique is being used to characterize the immune glycome.

Table 1: Advantages and disadvantages of mass spectrometry in glycan analysis

Advantages	Disadvantages
<ul style="list-style-type: none"> • Applicable for small sample amounts (lower fmol range) • Useful for complex glycan mixtures (generation of a further analysis dimension). • Attachment sides can be analysed by tandem MS experiments (side specific glycan analysis). • Glycan sequencing by tandem MS experiments. 	<ul style="list-style-type: none"> • Destructive method. • Need of a proper experimental design.

Arrays

Lectin and antibody arrays provide high-throughput screening of many samples containing glycans. This method uses either naturally occurring lectins or artificial monoclonal antibodies, where both are immobilized on a certain chip and incubated with a fluorescent glycoprotein sample.

Glycan arrays, like that offered by the Consortium for Functional Glycomics and Z Biotech LLC, contain carbohydrate compounds that can be screened with lectins or antibodies to define carbohydrate specificity and identify ligands.

Metabolic and covalent labeling of glycans

Metabolic labeling of glycans can be used as a way to detect glycan structures. A well known strategy involves the use of azide-labeled sugars which can be reacted using the Staudinger ligation. This method has been used for in vitro and in vivo imaging of glycans.

Tools for glycoproteins

X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy for complete structural analysis of complex glycans is a difficult and complex field. However, the structure of the binding site of numerous lectins, enzymes and other carbohydrate-binding proteins has revealed a wide variety of the structural basis for glycome function. The purity of test samples have been obtained through chromatography (affinity chromatography etc.) and analytical electrophoresis (PAGE (polyacrylamide electrophoresis), capillary electrophoresis, affinity electrophoresis, etc.).

Software and databases

There are several on-line software and databases available for glycomic research. This includes:

- GlycomeDB
- UniCarb-DB

Nutrigenomics

Nutrigenomics is a branch of nutritional genomics and is the study of the effects of foods and food constituents on gene expression. This means that nutrigenomics is research focusing on identifying and understanding molecular-level interaction between nutrients and other dietary bioactives with the genome. Nutrigenomics has also been described by the influence of genetic variation on nutrition, by correlating gene expression or SNPs with a nutrient's absorption, metabolism, elimination or biological effects. By doing so, nutrigenomics aims to develop rational means to optimise nutrition with respect to the subject's genotype.

By determining the mechanism of the effects of nutrients or the effects of a nutritional regime, nutrigenomics tries to define the causality or relationship between these specific nutrients and specific nutrient regimes (diets) on human health. Nutrigenomics has been associated with the idea of personalized nutrition based on genotype. While there is hope that nutrigenomics will ultimately enable such personalised dietary advice, it is a science still in its infancy and its contribution to public health over the next decade is thought to be major.^[3] Whilst nutrigenomics is aimed at developing an understanding of how the whole body responds to a food via systems biology, research into the effect of a single gene/single food compound relationships is known as nutrigenetics.

Definitions[edit]

In a *Nature Reviews Genetics* paper, nutrigenomics is defined as an emerging field of research that expands upon the existing field of nutritional science using genomic based data. Certain advances in the field such as microarrays, and high throughput sequencing allow for expansive analysis of the genome and in-vivo experiments in knockout mice are major sources of genomic based data. This type of genomic data collection can be applied to view the effects that certain nutrients or foods may have on large portions or different locales of the genome rather than one specific location.

Nutrigenomics is also defined as a field that examines "effect of nutrients on genome, proteome, metabolome and explains the relationship between these specific nutrients and nutrient-regimes on human health". In other words, a nutrigenomics approach is a holistic one that examines the effect of nutrients at all levels, from gene expression to metabolic pathways.

Background and preventive health[edit]

Nutritional science originally emerged as a field that studied individuals lacking certain nutrients and the subsequent effects, such as the disease scurvy which results from a lack of vitamin C. As other diseases closely related to diet,(but not deficiency) such as Obesity, became more prevalent, nutritional science expanded to cover these topics as well. Nutritional research typically focuses on preventative measure, trying to identify what nutrients or foods will raise or lower risks of diseases and damage to the human body.

Nutrigenomics emerged as a possible way to fix gaps in the current field of nutritional science. The development of technology to analyze the genome such as different types of sequencing and different microarrays suggest a new way to reinforce current theories or hypotheses. Existing information from genetic research directs emerging research in nutrigenomics. Individuals within the same population or even the same family have genetic variability. There is a lack of consistent relationships between certain foods and nutrients and increased disease risk, most likely due to this type of variation. Nutrigenomics is highly personalized because it looks at biomarkers within each individual. One group of researchers suggest that current technology can be used to build an ideal diet/intake of certain nutrients, or a 'nutriome.'A 'nutriome' would ensure proper function of all pathways involved in genome maintenance.

Research has already provided evidence identifying potential genetic origins of metabolic disorders or compromised phenotypes.^[5] Disorders that scientists previously thought to be heritable, can be identified as genetic disorders with set pathological effects. For example, Prader-Willi syndrome, a disease whose most distinguishing factor is insatiable appetite, has been specifically linked to an epigenetic pattern in which the paternal copy in the chromosomal region is erroneously deleted, and the maternal loci is inactivated by over methylation.Yet, although certain disorders may be linked to certain single nucleotide polymorphisms (SNPs) or other localized patterns, variation within a population may yield many more polymorphisms.Each may have a negligible effect by itself, yet the cumulative effects may be significant.Now, with advances that have been made, these small changes and additive effects are possible to study. Small epigenetic changes such as methylation patterns or phosphorylation can be determined.

Rationale and aims

Cell signaling is an important component of regulation of gene expression and metabolism, relying on both internal and external signals to ensure the body is maintaining homeostasis.

Individual nutrients can each be considered signals, with the summation of their effects being the diet.^[5] The effort of nutrigenomics is to identify this "dietary signature", or pattern of effects ranging from effects at the cellular level to entire body systems.^[5] However it is often hard to monitor the diet of an individual, and current protocols should be improved. The desired outcome from this type of research is to identify genetic factors for chronic diseases and conditions, whether it be a certain gene itself or an epigenetic marker, and how foods influence it. Nutrigenomics looks mainly to be a way of identifying individuals predisposed for conditions and preventing onset. First, genes with regulation influenced must be identified, and then more focused studies may emerge.

In addition, nutrigenomics also looks to identify certain compounds that are bioactive, and other foods that are of particular benefit to health. This knowledge can be personalized to produce specific diet plans and functional foods to both prevent predisposed conditions and maximize health.^[5]

Application

Anti-aging

Aging of cells occur because of the accumulation of excess free radicals formed due to the lack of proper nutrition to the cells and external factors like UV rays, pollution, stress, food, etc. DNA analysis is instrumental in identifying the right concoction of nutrients needed to eliminate the excess free radicals present in the cell.

The science of nutrigenomics studies the interaction between dietary components of food and genes.^[12] Scientific advances have now made it possible to apply nutrigenomics in the field of anti ageing and customize nutritional solutions in the form of supplements to meet the optimal nutrition required by the body to prevent aging of cells by the formation of excess free radicals.

Obesity

Obesity is one of the most widely studied topics in nutrigenomics. Due to genetic variations among individuals, each person could respond to diet differently. By exploring the interaction between dietary pattern and genetic factors, nutrigenomics aim to suggest prevention measures and/ or treatment to obesity via personal nutrition.

There are studies suggesting genetic factors account for a fair proportion of inter-individual BMI (Body Mass Index).^[15] Among different types of genetic variation between humans, SNPs are suggested to be the most important marker for the study of nutrigenomics. Multiple studies have found association between SNPs and obesity. One of the most well known obesity associating gene is the FTO gene. Among studied individuals, it was found

that those with AA genotype showed a higher BMI compared those with TT genotype when having high fat or low carbohydrate dietary intake.

The APO B SNP rs512535 is another obesity related variation. It was found that the A/G heterozygous genotype was found to have association with obesity (in terms of BMI and waist circumference). The same study also found that for individuals with habitual high fat diet (>35% of energy intaken), individuals with GG homozygotes genotype showed higher BMI compared to AA allele carriers. However, this difference is not found in low fat consuming group (<35% of energy intaken).

Besides the FTO genes and APO B, SNPs in various genes such as MC4R, SH2B1, MTCH2, SEC16B etc. have been found to be associated with obesity.^[15] Although many of these genetic variations are found in populations all over the world, there are also variations unique to certain races or populations.

Cancer[edit]

Nutrigenomics may be able to supplement current oncology. There is a wealth of information about processes that occur within genome maintenance that prevent cell abnormalities linked to cancer and certain nutrients that play a role as cofactors. Genome damage caused by micronutrient deficiency may be just as severe as damage owed to exposure to certain environmental carcinogens.^[8] If these micronutrients can be identified, with concrete evidence, the risk for cancer in some individuals could be significantly reduced. One such micronutrient may be folate. In one experiment, folate was given to cells in different concentrations and those with less folate exhibited as much damage to their chromosomes as they would have exhibited with a heavy amount of radiation.

Nutrigenomics can be used to develop new, alternative treatments that target the altered cancer cell metabolism. The alternative way of energy production in cancer cell metabolism, the Warburg effect, in which glycolysis and lactic acid fermentation are the main means of energy production opposed to oxidative reduction. Certain nutrients may provide ways to starve or inhibit this type of metabolism. Polyunsaturated fatty acids (PUFA) which affect gene expression related to inflammation and other nutrients that have displayed potential in repressing cancer cell metabolism. Another practical application of nutrigenomics to cancer may be identifying nutrient that is a cofactor of a compromised pathway where consuming a surplus of could potentially reduce the compromised pathway's negative consequences. A nutrigenomics approach could provide a safe, holistic model to mitigate tumor growth in place of existing cancer treatments that often have harsh side effects and are not always effective.

Companies involved

Companies across the Globe are currently involved in Nutrigenomics. Solutions such as genetic diet plans and exercise schedule have helped sportsmen perform even better. In India sportsmen like Sushil Kumar have taken advantage of Nutrigenomics. Today, even film stars have started following such plans. In the movie Dangal, Aamir Khan got his genes tested. Across the globe, there are companies who are not only working towards personalized nutrition but also personalized medicine. Few companies like DNAfit in London, 23andMe in the United States, Genecorp in India and HiMyDnain HongKong.

Ethics

To put nutrigenomics into practice, genetic testing is required as the test results act as the reference for diagnosis. Genetic testing has been met with many concerns surrounding ethics and regulations. These concerns inherently become a part of, if not augmented by Nutrigenomics, a field that looks to provide highly personalized information.

Consent

One of the major concerns regarding genetic tests would be privacy issue. To perform any type of genetic testing, consent is need directly by the individual who provides the sample. However, if an individual has results that indirectly tie family members to it, by identifying information about a genetic predisposition or condition, information about that family member has been inadvertently revealed. Thus, this type of genetic testing would require consent from a network of individuals. For some sets of the population such as mentally impaired adult or children, it is not possible to obtain direct consent. 'The best interest' of the patient must be determined by close family members, care takers and professionals, leaving room for discrepancy. Tissue samples obtained from patients, particularly those who are deceased are also a source of controversy. There is no established ethical code to suggest if data from these patients should be allowed to be published, or if they should remain only as sources of validation for lab techniques. There also exists no regulation for releasing information about heritable condition to family members. The stances on how to approach these situations are arbitrary and regulation provides few guidelines to direct them.

Distribution of tests

As the subject is recently commercialized by companies which sell direct to customer (DTC) genetic tests, as well as being applied by related professionals (such as dietetic practitioners), there has been increased awareness in the use of this information.

Validity

Nutrigenomics is still a new field. There are no set guidelines on how to interpret data from genetic testing. Without a validated way to produce accurate results, there exist concern about how valid results produced are. The Government Accountability Office (GAO) attempted to check the validity of numerous DTC tests by sending out information and samples of sham identities. The information they received was varied and not medically verified, and two companies tried to market general supplements as 'individualized'. The GAO study was also rudimentary, without taking into concern that differing environmental factors may affect results.

One suggestion to try and minimize fraud is to channel distribution of genetic testing to healthcare professionals. American College of Medical Genetics (ACMG) has taken a stance that healthcare professionals should be involved for proper implementation of information from genetic testing. Healthcare professionals are not necessarily qualified to properly interpret and distribute this information as it is not currently required that they have an in-depth knowledge of genetics. There are a sheer 45 genetic residencies in the US, with a low number of individuals who have completed training per year.^[24] Practitioners often focus on acute medical conditions and do not spend much of their time making health recommendations to each patient. It is suggested that nutritionists and genetic counselors may be the best choice to ensure proper distribution of genetic tests' results.

Privacy

One of the major concerns regarding genetic tests would be privacy issue. There are concerns on who has the right to have access to test results. Abuse of these tests could result in discrimination. For example, genetic information might be used by insurance companies to risk rate their clients or assess how likely their clients are to be costly. Other examples of privacy concerns include disclosure to the workplace that may led to discrimination in employment. Social concerns exist as certain conditions may be stigmatized by the general population.

Epigenomics

Epigenomics is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. The field is analogous to genomics and proteomics, which are the study of the genome and proteome of a cell. Epigenetic modifications are reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence. Epigenomic maintenance is a continuous process and plays an important role in stability of eukaryotic genomes by taking part in crucial biological mechanisms like DNA repair. Plant flavones are said to be inhibiting epigenomic marks that cause cancers. Two of the most characterized epigenetic modifications are DNA methylation and histone modification. Epigenetic modifications play an important role in gene expression and regulation, and are involved in numerous cellular processes such as in differentiation/development and tumorigenesis. The study of epigenetics on a global level has been made possible only recently through the adaptation of genomic high-throughput assays.

Introduction to Epigenetics

The mechanisms governing phenotypic plasticity, or the capacity of a cell to change its state in response to stimuli, have long been the subject of research (Phenotypic plasticity 1). The traditional central dogma of biology states that the DNA of a cell is transcribed to RNA, which is translated to proteins, which perform cellular processes and functions.^[10] A paradox exists, however, in that cells exhibit diverse responses to varying stimuli and that cells sharing identical sets of DNA such as in multicellular organisms can have a variety of distinct functions and phenotypes. Classical views have attributed phenotypic variation to differences in primary DNA structure, be it through aberrant mutation or an inherited sequence allele.^[12] However, while this did explain some aspects of variation, it does not explain how tightly coordinated and regulated cellular responses, such as differentiation, are carried out.

A more likely source of cellular plasticity is through the Regulation of gene expression, such that while two cells may have near identical DNA, the differential expression of certain genes results in variation. Research has shown that cells are capable of regulating gene expression at several stages: mRNA transcription, processing and transportation as well as in protein translation, post-translational processing and degradation. Regulatory proteins that bind to DNA, RNA, and/or proteins are key effectors in these

processes and function by positively or negatively regulating specific protein level and function in a cell.^[13] And, while DNA binding transcription factors provide a mechanism for specific control of cellular responses, a model where DNA binding transcription factors are the sole regulators of gene activity is also unlikely. For example, in a study of Somatic-cell nuclear transfer, it was demonstrated that stable features of differentiation remain after the nucleus is transferred to a new cellular environment, suggesting that a stable and heritable mechanism of gene regulation was involved in the maintenance of the differentiated state in the absence of the DNA binding transcription factors.^[11]

With the finding that DNA methylation and histone modifications are stable, heritable, and also reversible processes that influence gene expression without altering DNA primary structure, a mechanism for the observed variability in cell gene expression was provided.^[12] These modifications were termed epigenetic, from epi “on top of” the genetic material “DNA” (Epigenetics 1). The mechanisms governing epigenetic modifications are complex, but through the advent of high-throughput sequencing technology they are now becoming better understood.^[12]

Epigenetics

Genomic modifications that alter gene expression that cannot be attributed to modification of the primary DNA sequence and that are heritable mitotically and meiotically are classified as epigenetic modifications. DNA methylation and histone modification are among the best characterized epigenetic processes.^[3]

DNA methylation

The first epigenetic modification to be characterized in depth was DNA methylation. As its name implies, DNA methylation is the process by which a methyl group is added to DNA. The enzymes responsible for catalyzing this reaction are the DNA methyltransferases (DNMTs). While DNA methylation is stable and heritable, it can be reversed by an antagonistic group of enzymes known as DNA de-methylases. In eukaryotes, methylation is most commonly found on the carbon 5 position of cytosine residues (5mC) adjacent to guanine, termed CpGdinucleotides.^{[9][14]}

DNA methylation patterns vary greatly between species and even within the same organism. The usage of methylation among animals is quite different; with vertebrates exhibiting the highest levels of 5mC and invertebrates more moderate levels of 5mC. Some organisms like *Caenorhabditiselegans* have not been demonstrated to have 5mC nor a conventional

DNA methyltransferase; this would suggest that other mechanisms other than DNA methylation are also involved.^[11]

Within an organism, DNA methylation levels can also vary throughout development and by region. For example, in mouse primordial germ cells, a genome wide de-methylation even occurs; by implantation stage, methylation levels return to their prior somatic levels.^[11] When DNA methylation occurs at promoter regions, the sites of transcription initiation, it has the effect of repressing gene expression. This is in contrast to unmethylated promoter regions which are associated with actively expressed genes.^[9]

The mechanism by which DNA methylation represses gene expression is a multi-step process. The distinction between methylated and unmethylated cytosine residues is carried out by specific DNA-binding proteins. Binding of these proteins recruit histone deacetylases (HDACs) enzyme which initiate chromatin remodeling such that the DNA becoming less accessible to transcriptional machinery, such as RNA polymerase, effectively repressing gene expression.^[15]

Histone Modification

In eukaryotes, genomic DNA is coiled into protein-DNA complexes called chromatin. Histones, which are the most prevalent type of protein found in chromatin, function to condense the DNA; the net positive charge on histones facilitates their bonding with DNA, which is negatively charged. The basic and repeating units of chromatin, nucleosomes, consist of an octamer of histone proteins (H2A, H2B, H3 and H4) and a 146 bp length of DNA wrapped around it. Nucleosomes and the DNA connecting form a 10 nm diameter chromatin fiber, which can be further condensed.^{[16][17]}

Chromatin packaging of DNA varies depending on the cell cycle stage and by local DNA region. The degree to which chromatin is condensed is associated with a certain transcriptional state. Unpackaged or loose chromatin is more transcriptionally active than tightly packaged chromatin because it is more accessible to transcriptional machinery. By remodeling chromatin structure and changing the density of DNA packaging, gene expression can thus be modulated.

Chromatin remodeling occurs via post-translational modifications of the N-terminal tails of core histone proteins.^[19] The collective set of histone modifications in a given cell is known as the histone code. Many different types of histone modification are known, including: acetylation, methylation, phosphorylation, ubiquitination, SUMOylation, ADP-ribosylation, deamination and proline isomerization; acetylation, methylation, phosphorylation and ubiquitination have been implicated in gene activation whereas

methylation, ubiquitination, SUMOylation, deamination and proline isomerization have been implicated in gene repression. Note that several modification types including methylation, phosphorylation and ubiquitination can be associated with different transcriptional states depending on the specific amino acid on the histone being modified. Furthermore, the DNA region where histone modification occurs can also elicit different effects; an example being methylation of the 3rd core histone at lysine residue 36 (H3K36). When H3K36 occurs in the coding sections of a gene, it is associated with gene activation but the opposite is found when it is within the promoter region.^[17]

Histone modifications regulate gene expression by two mechanisms: by disruption of the contact between nucleosomes and by recruiting chromatin remodeling ATPases. An example of the first mechanism occurs during the acetylation of lysine terminal tail amino acids, which is catalyzed by histone acetyltransferases (HATs). HATs are part of a multiprotein complex that is recruited to chromatin when activators bind to DNA binding sites. Acetylation effectively neutralizes the basic charge on lysine, which was involved in stabilizing chromatin through its affinity for negatively charged DNA. Acetylated histones therefore favor the dissociation of nucleosomes and thus unwinding of chromatin can occur. Under a loose chromatin state, DNA is more accessible to transcriptional machinery and thus expression is activated. The process can be reversed through removal of histone acetyl groups by deacetylases. The second process involves the recruitment of chromatin remodeling complexes by the binding of activator molecules to corresponding enhancer regions. The nucleosome remodeling complexes reposition nucleosomes by several mechanisms, enabling or disabling accessibility of transcriptional machinery to DNA. The SWI/SNF protein complex in yeast is one example of a chromatin remodeling complex that regulates the expression of many genes through chromatin remodeling.^{[17][20]}

Relation to other genomic fields

Epigenomics shares many commonalities with other genomics fields, in both methodology and in its abstract purpose. Epigenomics seeks to identify and characterize epigenetic modifications on a global level, similar to the study of the complete set of DNA in genomics or the complete set of proteins in a cell in proteomics.^{[1][2]} The logic behind performing epigenetic analysis on a global level is that inferences can be made about epigenetic modifications, which might not otherwise be possible through analysis of specific loci.^{[16][1]} As in the other genomics fields, epigenomics relies heavily on bioinformatics, which combines the disciplines of biology, mathematics and computer science. However

while epigenetic modifications had been known and studied for decades, it is through these advancements in bioinformatics technology that have allowed analyses on a global scale. Many current techniques still draw on older methods, often adapting them to genomic assays as is described in the next section.

Epigenomics Methods

Histone modification assays

The cellular processes of transcription, DNA replication and DNA repair involve the interaction between genomic DNA and nuclear proteins. It had been known that certain regions within chromatin were extremely susceptible to DNase I digestion, which cleaves DNA in a low sequence specificity manner. Such hypersensitive sites were thought to be transcriptionally active regions, as evidenced by their association with RNA polymerase and topoisomerases I and II.

It is now known that sensitivity to DNase I regions correspond to regions of chromatin with loose DNA-histone association. Hypersensitive sites most often represent promoters regions, which require for DNA to be accessible for DNA binding transcriptional machinery to function.^[23]

ChIP-Chip and ChIP-Seq

Histone modification was first detected on a genome wide level through the coupling of chromatin immunoprecipitation (ChIP) technology with DNA microarrays, termed ChIP-Chip.^[16] However instead of isolating a DNA-binding transcription factor or enhancer protein through chromatin immunoprecipitation, the proteins of interest are the modified histones themselves. First, histones are cross-linked to DNA in vivo through light chemical treatment (e.g., formaldehyde). The cells are next lysed, allowing for the chromatin to be extracted and fragmented, either by sonication or treatment with a non-specific restriction enzyme (e.g., micrococcal nuclease). Modification-specific antibodies in turn, are used to immunoprecipitate the DNA-histone complexes.^[17] Following immunoprecipitation, the DNA is purified from the histones, amplified via PCR and labeled with a fluorescent tag (e.g., Cy5, Cy3). The final step involves hybridization of labeled DNA, both immunoprecipitated DNA and non-immunoprecipitated onto a microarray containing immobilized gDNA. Analysis of the relative signal intensity allows the sites of histone modification to be determined.^{[24][25]}

ChIP-chip was used extensively to characterize the global histone modification patterns of yeast. From these studies, inferences on the function of histone modifications were made; that transcriptional activation or repression was associated with certain histone modifications

and by region. While this method was effective providing near full coverage of the yeast epigenome, its use in larger genomes such as humans is limited.^{[16][17]}

In order to study histone modifications on a truly genome level, other high-throughput methods were coupled with the chromatin immunoprecipitation, namely: SAGE: serial analysis of gene expression (ChIP-SAGE), PET: paired end ditag sequencing (ChIP-PET) and more recently, next-generation sequencing (ChIP-Seq). ChIP-seq follows the same protocol for chromatin immunoprecipitation but instead of amplification of purified DNA and hybridization to a microarray, the DNA fragments are directly sequenced using next generation parallel re-sequencing. It has proven to be an effective method for analyzing the global histone modification patterns and protein target sites, providing higher resolution than previous methods.^{[16][24]}

DNA Methylation assays

Techniques for characterizing primary DNA sequences could not be directly applied to methylation assays. For example, when DNA was amplified in PCR or bacterial cloning techniques, the methylation pattern was not copied and thus the information lost. The DNA hybridization technique used in DNA assays, in which radioactive probes were used to map and identify DNA sequences, could not be used to distinguish between methylated and non-methylated DNA.^{[26][9]}

Restriction endonuclease based methods

Non genome-wide approaches

The earliest methylation detection assays used methylation modification sensitive restriction endonucleases. Genomic DNA was digested with both methylation-sensitive and insensitive restriction enzymes recognizing the same restriction site. The idea being that whenever the site was methylated, only the methylation insensitive enzyme could cleave at that position. By comparing restriction fragment sizes generated from the methylation-sensitive enzyme to those of the methylation-insensitive enzyme, it was possible to determine the methylation pattern of the region. This analysis step was done by amplifying the restriction fragments via PCR, separating them through gel electrophoresis and analyzing them via southern blot with probes for the region of interest.^{[26][9]}

This technique was used to compare the DNA methylation modification patterns in the human adult and hemoglobin gene loci. Different regions of the gene (gamma delta beta globin) were known to be expressed at different stages of development.^[27] Consistent with a role of DNA methylation in gene repression, regions that were associated with high levels of DNA methylation were not actively expressed.^[28]

This method was limited not suitable for studies on the global methylation pattern, or ‘methylome’. Even within specific loci it was not fully representative of the true methylation pattern as only those restriction sites with corresponding methylation sensitive and insensitive restriction assays could provide useful information. Further complications could arise when incomplete digestion of DNA by restriction enzymes generated false negative results.^[9]

Genome wide approaches

DNA methylation profiling on a large scale was first made possible through the Restriction Landmark Genome Scanning (RLGS) technique. Like the locus-specific DNA methylation assay, the technique identified methylated DNA via its digestion methylation sensitive enzymes. However it was the use of two-dimensional gel electrophoresis that allowed be characterized on a broader scale.

However it was not until the advent of microarray and next generation sequencing technology when truly high resolution and genome-wide DNA methylation became possible.^[12] As with RLGS, the endonuclease component is retained in the method but it is coupled to new technologies. One such approach is the differential methylation hybridization (DMH), in which one set of genomic DNA is digested with methylation-sensitive restriction enzymes and a parallel set of DNA is not digested. Both sets of DNA are subsequently amplified and each labelled with fluorescent dyes and used in two-colour array hybridization. The level of DNA methylation at a given loci is determined by the relative intensity ratios of the two dyes. Adaptation of next generation sequencing to DNA methylation assay provides several advantages over array hybridization. Sequence-based technology provides higher resolution to allele specific DNA methylation, can be performed on larger genomes, and does not require creation of DNA microarrays which require adjustments based on CpG density to properly function.

Bisulfite sequencing

Bisulfite sequencing relies on chemical conversion of unmethylated cytosines exclusively, such that they can be identified through standard DNA sequencing techniques. Sodium bisulfate and alkaline treatment does this by converting unmethylated cytosine residues into uracil while leaving methylated cytosine unaltered. Subsequent amplification and sequencing of untreated DNA and sodium bisulphite treated DNA allows for methylated sites to be identified. Bisulfite sequencing, like the traditional restriction based methods, was historically limited to methylation patterns of specific gene loci, until whole genome sequencing technologies became available. However, unlike traditional restriction based methods, bisulfite sequencing provided resolution on a nucleotide level.

Limitations of the bisulfite technique include the incomplete conversion of cytosine to uracil, which is a source of false positives. Further, bisulfite treatment also causes DNA degradation and requires an additional purification step to remove the sodium bisulfite.^[9]

Next-generation sequencing is well suited in complementing bisulfite sequencing in genome-wide methylation analysis. While this now allows for methylation pattern to be determined on the highest resolution possible, on the single nucleotide level, challenges still remain in the assembly step because of reduced sequence complexity in bisulphite treated DNA. Increases in read length seek to address this challenge, allowing for whole genome shotgun bisulphite sequencing (WGBS) to be performed. The WGBS approach using an Illumina Genome Analyzer platform and has already been implemented in *Arabidopsis thaliana*.^[9]

Direct Detection

Polymerase sensitivity in single molecule real time sequencing made it possible for scientists to directly detect epigenetic marks such as methylation as the polymerase moves along the DNA molecule being sequenced.^[29] Several projects have demonstrated the ability to collect genome-wide epigenetic data in bacteria.

Theoretical modeling approaches

First mathematical models for different nucleosome states affecting gene expression were introduced in 1980s [ref]. Later, this idea was almost forgotten, until the experimental evidence has indicated a possible role of covalent histone modifications as an epigenetic code. In the next several years, high-throughput data have indeed uncovered the abundance of epigenetic modifications and their relation to chromatin functioning which has motivated new theoretical models for the appearance, maintaining and changing these patterns. These models are usually formulated in the frame of one-dimensional lattice approaches