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 BIOINFORMATICS

UNIT – 1- SBIA1301 – Molecular Biology and Genomics

DNA

Deoxyribonucleic acid (/diːˈɒksɪˌraɪboʊnjuːˌkliːɪk, -ˌkleɪ-/ (listen);[1] DNA) is a molecule composed of two polynucleotide chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. DNA and ribonucleic acid (RNA) are nucleic acids. Alongside proteins, lipids and complex carbohydrates (polysaccharides), nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life.

The two DNA strands are known as polynucleotides as they are composed of simpler monomeric units called nucleotides.[2][3] Each nucleotide is composed of one of four nitrogen-containing nucleobases (cytosine [C], guanine [G], adenine [A] or thymine [T]), a sugar called deoxyribose, and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds (known as the phospho-diester linkage) between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. The nitrogenous bases of the two separate polynucleotide strands are bound together, according to base pairing rules (A with T and C with G), with hydrogen bonds to make double-stranded DNA. The complementary nitrogenous bases are divided into two groups, pyrimidines and purines. In DNA, the pyrimidines are thymine and cytosine; the purines are adenine and guanine.

Both strands of double-stranded DNA store the same biological information. This information is replicated as and when the two strands separate. A large part of DNA (more than 98% for humans) is non-coding, meaning that these sections do not serve as patterns for protein sequences. The two strands of DNA run in opposite directions to each other and are thus antiparallel. Attached to each sugar is one of four types of nucleobases (informally, *bases*). It is the sequence of these four nucleobases along the backbone that encodes genetic information. RNA strands are created using DNA strands as a template in a process called transcription, where DNA bases are exchanged for their corresponding bases except in the case of thymine (T), for which RNA substitutes uracil (U).[4] Under the genetic code, these RNA strands specify the sequence of amino acids within proteins in a process called translation.

Within eukaryotic cells, DNA is organized into long structures called chromosomes. Before typical cell division, these chromosomes are duplicated in the process of DNA replication, providing a complete set of chromosomes for each daughter cell. Eukaryotic organisms (animals, plants, fungi and protists) store most of their DNA inside the cell nucleus as nuclear DNA, and some in the mitochondria as mitochondrial DNA or in chloroplasts as chloroplast DNA.[5] In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm, in circular chromosomes. Within eukaryotic chromosomes, chromatin proteins, such as histones, compact and organize DNA. These compacting structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

DNA was first isolated by Friedrich Miescher in 1869. Its molecular structure was first identified by Francis Crick and James Watson at the Cavendish Laboratory within the University of Cambridge in 1953, whose model-building efforts were guided by X-ray diffraction data

acquired by Raymond Gosling, who was a post-graduate student of Rosalind Franklin at King's College London. DNA is used by researchers as a molecular tool to explore physical laws and theories, such as the ergodic theorem and the theory of elasticity. The unique material properties of DNA have made it an attractive molecule for material scientists and engineers interested in micro- and nano-fabrication. Among notable advances in this field are DNA origami and DNA-based hybrid materials.

Sense and antisense

A DNA sequence is called a "sense" sequence if it is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands can contain both sense and antisense sequences). In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear.[33] One proposal is that antisense RNAs are involved in regulating gene expression through RNA-RNA base pairing.

A few DNA sequences in prokaryotes and eukaryotes, and more in plasmids and viruses, blur the distinction between sense and antisense strands by having overlapping genes. In these cases, some DNA sequences do double duty, encoding one protein when read along one strand, and a second protein when read in the opposite direction along the other strand. In bacteria, this overlap may be involved in the regulation of gene transcription, while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome

Watson and Crick

In 1951, the then 23-year old biologist James Watson travelled from the United States to work with Francis Crick, an English physicist at the University of Cambridge. Crick was already using the process of X-ray crystallography to study the structure of protein molecules. Together, Watson and Crick used X-ray crystallography data, produced by Rosalind Franklin and Maurice Wilkins at King's College in London, to decipher DNA's structure.

This is what they already knew from the work of many scientists, about the DNA molecule:

1. DNA is made up of subunits which scientists called nucleotides.
2. Each nucleotide is made up of a sugar, a phosphate and a base.
3. There are 4 different bases in a DNA molecule:
   adenine (a purine)
   cytosine (a pyrimidine)
   guanine (a purine)
   thymine (a pyrimidine)
4. The number of purine bases equals the number of pyrimidine bases
5. The number of adenine bases equals the number of thymine bases

6. The number of guanine bases equals the number of cytosine bases
7. The basic structure of the DNA molecule is helical, with the bases being stacked on top of each other

Components of DNA

DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide". Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one letter abbreviations as shorthand for the four bases.

- A is for adenine
- G is for guanine
- C is for cytosine
- T is for thymine

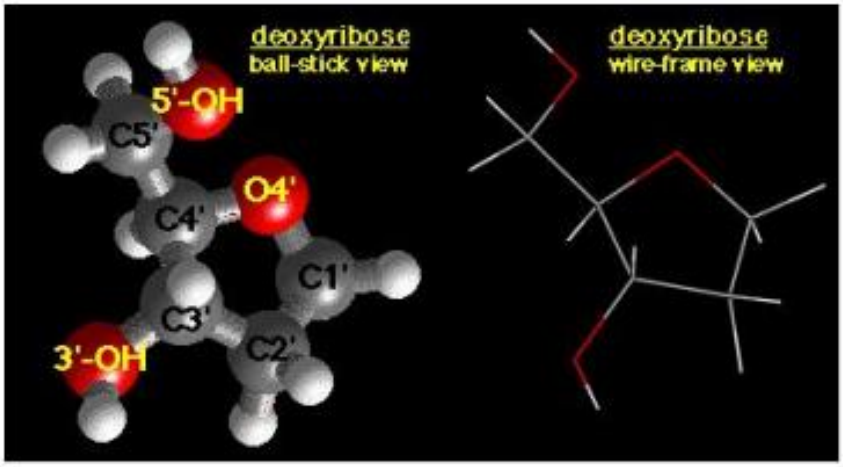
Purine Bases

Adenine and guanine are purines. Purines are the larger of the two types of bases found in DNA. Structures are shown below:

The 9 atoms that make up the fused rings (5 carbon, 4 nitrogen) are numbered 1-9. All ring atoms lie in the same plane.

Deoxyribose Sugar

The deoxyribose sugar of the DNA backbone has 5 carbons and 3 oxygens. The carbon atoms are numbered 1', 2', 3', 4', and 5' to distinguish from the numbering of the atoms of the purine and pyrmidine rings. The hydroxyl groups on the 5'- and 3'- carbons link to the phosphate groups to form the DNA backbone. Deoxyribose lacks an hydroxyl group at the 2'-position when compared to ribose, the sugar component of RNA.



Nucleosides

A nucleoside is one of the four DNA bases covalently attached to the C1' position of a sugar. The sugar in deoxynucleosides is 2'-deoxyribose. The sugar in ribonucleosides is ribose. Nucleosides differ from nucleotides in that they lack phosphate groups. The four different nucleosides of DNA are deoxyadenosine (dA), deoxyguanosine (dG), deoxycytosine (dC), and (deoxy)thymidine (dT, or T).

In dA and dG, there is an "N-glycoside" bond between the sugar C1' and N9 of the purine.

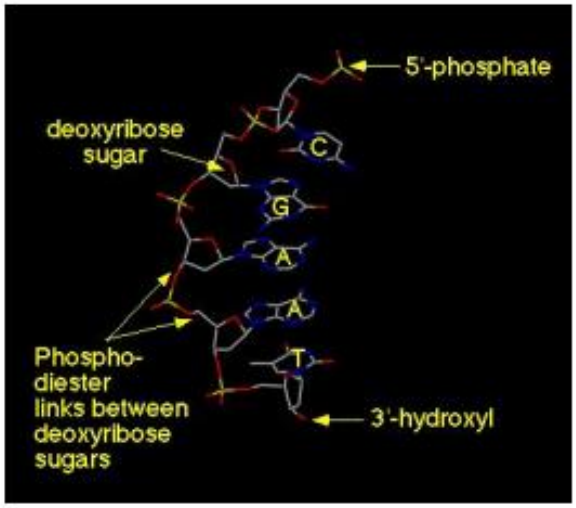
Nucleotides

A nucleotide is a nucleoside with one or more phosphate groups covalently attached to the 3'- and/or 5'-hydroxyl group(s).

DNA Backbone

The DNA backbone is a polymer with an alternating sugar-phosphate sequence. The deoxyribose sugars are joined at both the 3'-hydroxyl and 5'-hydroxyl groups to phosphate groups in ester links, also known as "phosphodiester" bonds.

Example of DNA Backbone: 5'-d (CGAAT)



Features of the 5'-d(CGAAT) structure:

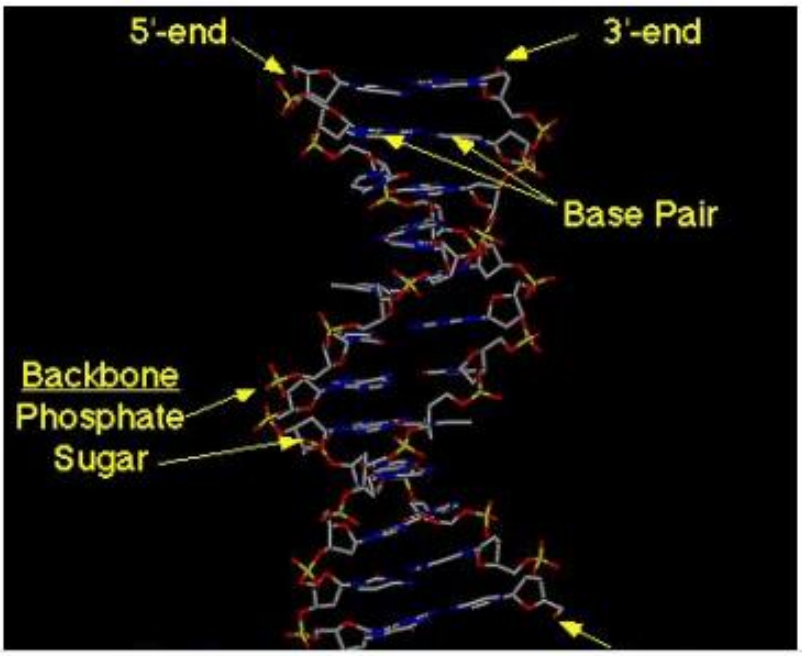
- Alternating backbone of deoxyribose and phosphodiester groups

- Chain has a direction (known as polarity), 5'- to 3'- from top to bottom
- Oxygens (red atoms) of phosphates are polar and negatively charged
- A, G, C, and T bases can extend away from chain, and stack atop each other
- Bases are hydrophobic

DNA Double Helix

DNA is a normally double stranded macromolecule. Two polynucleotide chains, held together by weak thermodynamic forces, form a DNA molecule.

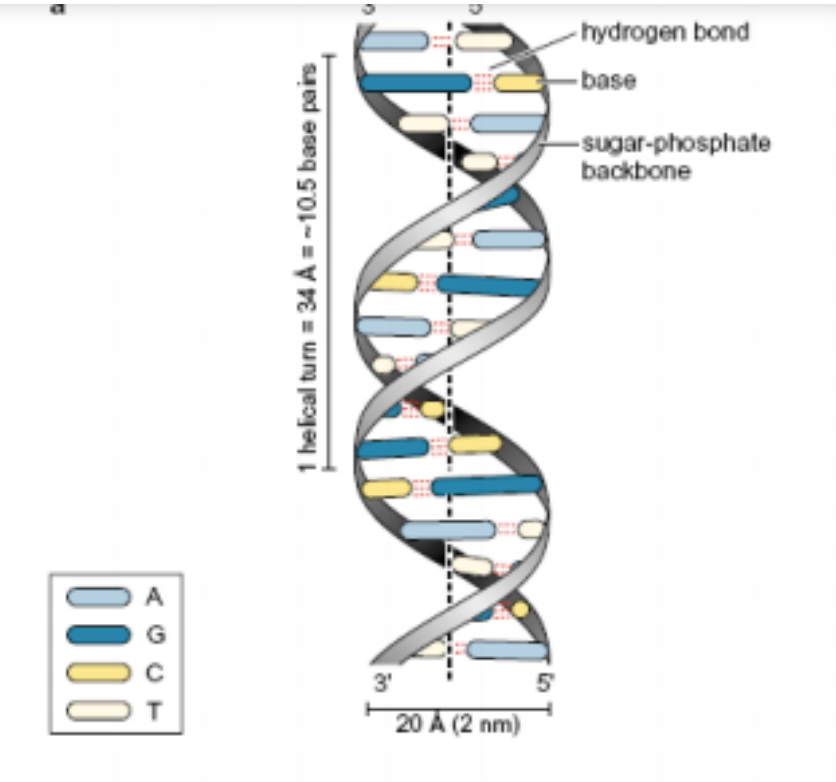
Structure of DNA Double Helix



Features of the DNA Double Helix

- Two DNA strands form a helical spiral, winding around a helix axis in a right-handed spiral
- The two polynucleotide chains run in opposite directions
- The sugar-phosphate backbones of the two DNA strands wind around the helix axis like the railing of a sprial staircase

The bases of the individual nucleotides are on the inside of the helix, stacked on top of each other like the steps of a spiral staircase.

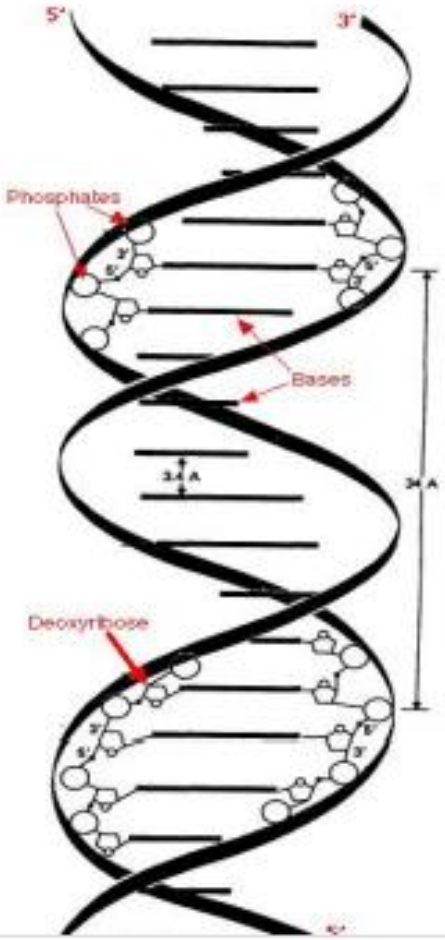


The Double Helix

The double helix of DNA has these features:

- It contains two polynucleotide strands wound around each other.
- The backbone of each consists of alternating deoxyribose and phosphate groups.
- The phosphate group bonded to the 5' carbon atom of one deoxyribose is covalently bonded to the 3' carbon of the next.
- The two strands are "antiparallel"; that is, one strand runs 5′ to 3′ while the other runs 3′ to 5′.
- The DNA strands are assembled in the 5′ to 3′ direction and, by convention, we "read" them the same way.
- The purine or pyrimidine attached to each deoxyribose projects in toward the axis of the helix.
- Each base forms hydrogen bonds with the one directly opposite it, forming **base pairs** (also called nucleotide pairs).

- 3.4 Å separate the planes in which adjacent base pairs are located.
- The double helix makes a complete turn in just over 10 nucleotide pairs, so each turn takes a little more (35.7 Å to be exact) than the 34 Å shown in the diagram.
- There is an average of 25 hydrogen bonds within each complete turn of the double helix providing a stability of binding about as strong as what a covalent bond would provide.
- The diameter of the helix is 20 Å.
- The helix can be virtually any length; when fully stretched, some DNA molecules are as much as 5 cm (2 inches!) long.
- The path taken by the two backbones forms a major (wider) groove (from "34 A" to the top of the arrow) and a minor (narrower) groove (the one below).



Nucleic acids (DNA and RNA) are the polymers i.e. long chain compounds. The molecular structure of DNA has two aspects

1) its chemical sub units and

2) the way in which these chemical sub units are arranged to form a long chain molecule.

The second aspect is very significant as the accepted DNA model should be such that it explains biochemically the various aspects (function) of a gene such as stability to metabolic and external agents, the capacity for replication (self duplication) the capacity to store vast hereditary information in coded form and the capacity to express the phenotypes they control.

FUNCTIONS OF DNA

DNA carries the genetic information of a cell and consists of thousands of genes. Each gene serves as a recipe on how to build a protein molecule. Proteins perform important tasks for the cell functions or serve as building blocks. The flow of information from the genes determines the protein composition and thereby the functions of the cell.

The DNA is situated in the nucleus, organized into chromosomes. Every cell must contain the genetic information and the DNA is therefore duplicated before a cell divides (**replication**). When proteins are needed, the corresponding genes are transcribed into RNA (**transcription**). The RNA is first processed so that non-coding parts are removed (**processing**) and is then transported out of the nucleus (**transport**). Outside the nucleus, the proteins are built based upon the code in the RNA (**translation**).

Types of DNA

DNA can be classified in various ways based on 1. number of base pair per turn. 2. coiling pattern, 3. location 4. structure, 5. nucleotide sequence and 6. number of strands.

1. Number of base per turn. Depending upon the nucleotide base per turn of the helix, tilt of the base pair and humidity of the sample, the DNA can be observed in four different forms namely A,B, C and D.

2. Coiling pattern. On the basis of coiling pattern of the helix DNA is of two types viz right handed and left handed. Most of the DNA molecules are right handed i.e. coiling of helix is in the right direction. It is also

called positive coiling. All the four forms of DNA viz A, B, C and D are right handed. The Z DNA has left handed double helical structure. This DNA is considered to be associated with gene regulation.

3. Location. Based on the location in the cell DNA is of three types. Viz., chromosomal DNA cytoplasm DNA and promiscuous DNA. Chromosomal DNA is found in chromosomes. And are called as chromosomal DNA or nuclear DNA. Cytoplasmic DNA is found in the cytoplasm especially in mitochondria and chloroplasts. Such DNA plays an important role in cytoplasmic inheritance and has circular structure. Promiscuous DNA. Some DNA segments with common base sequence are found in the chloroplasts, mitochondria and nucleus. This suggests that some DNA sequences move from one organelle to other. Such DNA is referred to as promiscuous DNA.

RNA- Properties, Structure, Types and Functions

- RNA or ribonucleic acid is a polymer of nucleotides which is made up of a ribose sugar, a phosphate, and bases such as adenine, guanine, cytosine, and uracil.
- It is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression of genes.
-

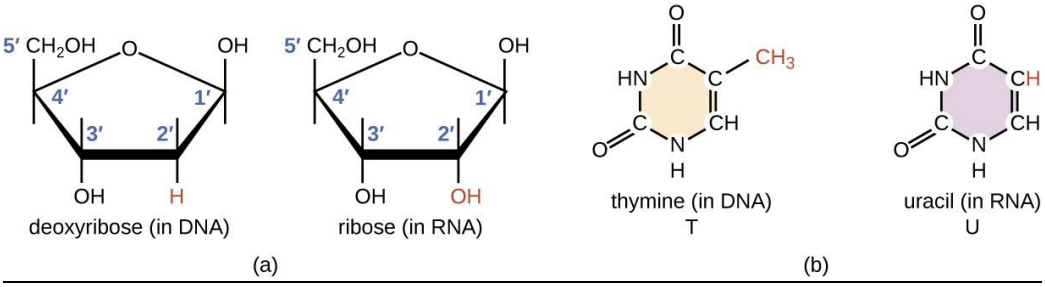
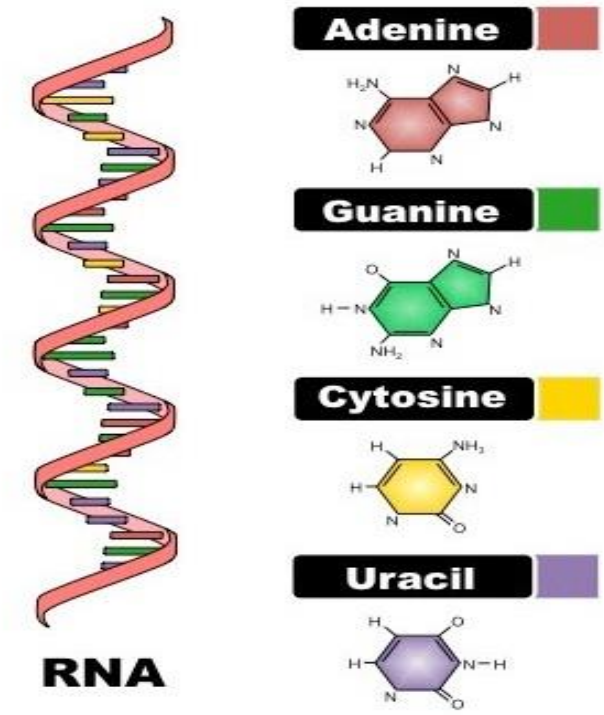


Figure: (a) Ribonucleotides contain the pentose sugar ribose instead of the deoxyribose found in deoxyribonucleotides. (b) RNA contains the pyrimidine uracil in place of thymine found in DNA.

RNA STRUCTURE



Like **DNA**, RNA is a long polymer consisting of nucleotides.

- RNA is a single-stranded helix.
- The strand has a 5′end (with a phosphate group) and a 3′end (with a hydroxyl group).
- It is composed of ribonucleotides.
- The ribonucleotides are linked together by 3′ –> 5′ phosphodiester bonds.
- The nitrogenous bases that compose the ribonucleotides include adenine, cytosine, uracil, and guanine.

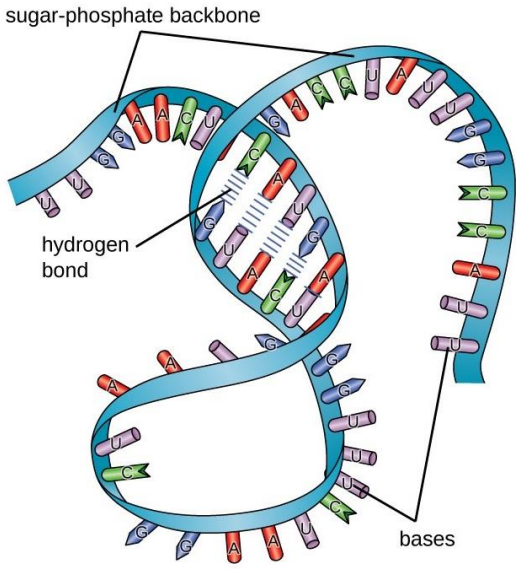
Thus, the difference in the structure of RNA from that of DNA include:

- The bases in RNA are adenine (abbreviated A), guanine (G), uracil (U) andcytosine (C).

Thus thymine in DNA is replaced by uracil in RNA, a different pyrimidine. However, like thymine, uracil can form base pairs with adenine.

- The sugar in RNA is ribose rather than deoxyribose as in DNA.
- The corresponding ribonucleosides are adenosine, guanosine, cytidine and uridine. The corresponding ribonucleotides are adenosine 5’-triphosphate (ATP), guanosine 5’-triphosphate (GTP), cytidine 5’-triphosphate (CTP) and uridine 5’-triphosphate (UTP).

RNA Secondary structure



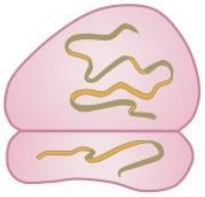
- Most RNA molecules are single-stranded but an RNA molecule may contain regions which can form complementary base pairing where the RNA strand loops back on itself.
- If so, the RNA will have some double-stranded regions.
- Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) exhibit substantial secondary structure, as do some messenger RNAs (mRNAs).

Types of RNA

In prokaryotes and eukaryotes, there are three main types of RNA.

- rRNA
- mRNA
- tRNA

Messenger RNA (mRNA)



Ribosomal RNA (rRNA)

Transfer RNA (tRNA)

Messenger RNA (mRNA)

- Accounts for about 5% of the total RNA in the cell.
- Most heterogeneous of the 3 types of RNA in terms of both base sequence and size.
- It carries the genetic code copied from the DNA during transcription in the form of triplets of nucleotides called codons.
- As part of post-transcriptional processing in eukaryotes, the 5′ end of mRNA is capped with a guanosine triphosphate nucleotide, which helps in mRNA recognition during translation or protein synthesis.
- Similarly, the 3′ end of an mRNA has a poly A tail or multiple adenylate residues added to it, which prevent enzymatic degradation of mRNA. Both 5′ and 3′ end of an mRNA imparts stability to the mRNA.

Function

mRNA transcribes the genetic code from DNA into a form that can be read and used to make proteins. mRNA carries genetic information from the nucleus to the cytoplasm of a cell.

Ribosomal RNA (rRNA)

- Found in the ribosomes and account for 80% of the total RNA present in the cell.
- Ribosomes consist of two major components: the small ribosomal subunits, which read the RNA, and the large subunits, which join amino acids to form a polypeptide chain. Each subunit comprises one or more ribosomal RNA (rRNA) molecules and a variety of ribosomal proteins (r-protein or rProtein).
- Different rRNAs present in the ribosomes include small rRNAs and large rRNAs, which denote their presence in the small and large subunits of the ribosome.
- rRNAs combine with proteins in the cytoplasm to form ribosomes, which act as the site of protein synthesis and has the enzymes needed for the process.
- These complex structures travel along the mRNA molecule during translation and facilitate the assembly of amino acids to form a polypeptide chain. They bind to tRNAs and other molecules that are crucial for protein synthesis.

Function

rRNA directs the translation of mRNA into proteins.

Transfer RNA (tRNA)

- tRNA is the smallest of the 3 types of RNA having about 75-95 nucleotides.
- tRNAs are an essential component of translation, where their main function is the transfer of amino acids during protein synthesis. Therefore they are called transfer RNAs.
- Each of the 20 amino acids has a specific tRNA that binds with it and transfers it to the growing polypeptide chain. tRNAs also act as adapters in the translation of the genetic sequence of mRNA into proteins. Therefore they are also called adapter molecules.

Structure of tRNA

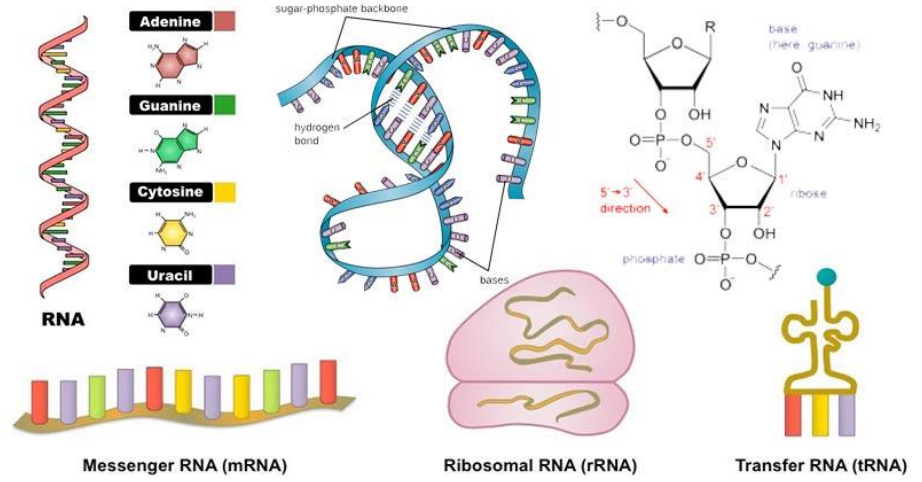
tRNAs have a clover leaf structure which is stabilized by strong hydrogen bonds between the nucleotides. Apart from the usual 4 bases, they normally contain some unusual bases mostly formed by methylation of the usual bases, for example, methyl guanine and methylcytosine.

- Three structural loops are formed via hydrogen bonding.
- The 3′ end serves as the amino acid attachment site.
- The center loop encompasses the anticodon.
- The anticodon is a three-base nucleotide sequence that binds to the mRNA codon.
- This interaction between codon and anticodon specifies the next amino acid to be added during protein synthesis.

Function

Transfer RNA brings or transfers amino acids to the ribosome that correspond to each three-nucleotide codon of rRNA. The amino acids then can be joined together and processed to make polypeptides and proteins.

Other Properties of RNA



- RNA forms in the nucleolus, and then moves to specialized regions of the cytoplasm depending on the type of RNA formed.
- RNA, containing a ribose sugar, is more reactive than DNA and is not stable in alkaline conditions. RNA's larger helical grooves mean it is more easily subject to attack by enzymes.
- RNA strands are continually made, broken down and reused.
- RNA is more resistant to damage from UV light than DNA.
- RNA's mutation rate is relatively higher.
- Unusual bases may be present.
- The number of RNA may differ from cell to cell.
- Rate of renaturation after melting is quick.
- RNA is more versatile than DNA, capable of performing numerous, diverse tasks in an organism.

FUNCTIONS OF RNA

- RNA is a nucleic acid messenger between DNA and ribosomes.
- It serves as the genetic material in some organisms (viruses).
- Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals.
- Messenger RNA (mRNA) copies DNA in the nucleus and carries the info to the ribosomes (in cytoplasm).
- Ribosomal RNA (rRNA) makes up a large part of the ribosome; reads and decodes mRNA.
- Transfer RNA (tRNA) carries amino acids to the ribosome where they are joined to form proteins.
- Certain RNAs are able to catalyse chemical reactions such as cutting and ligating other RNA molecules, and the catalysis of peptide bond formation in the ribosome; these are known as ribozymes.

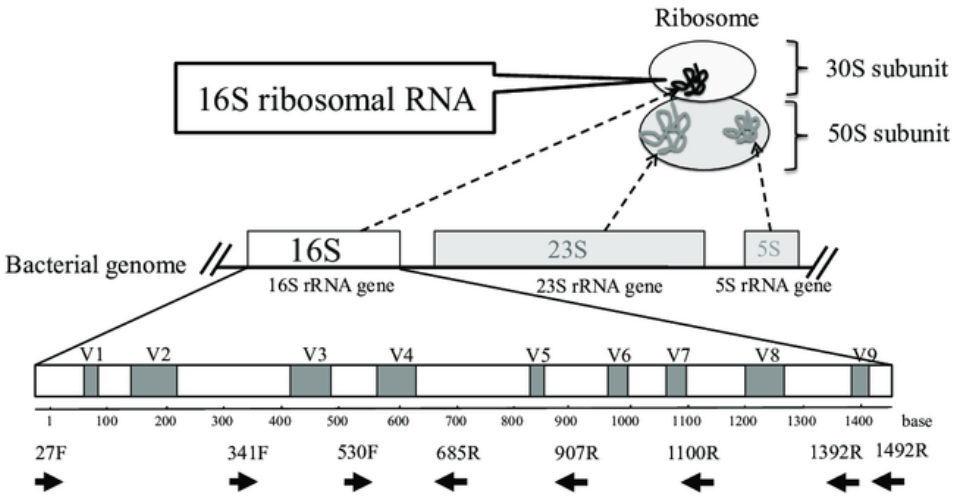
16srRNA

16S ribosomal RNA (or 16S rRNA) is the RNA component of the 30S subunit of a prokaryotic ribosome (SSU rRNA). It binds to the Shine-Dalgarno sequence and provides most of the SSU structure.

The genes coding for it are referred to as **16S rRNA gene** and are used in reconstructing phylogenies, due to the slow rates of evolution of this region of the gene.[2] Carl Woese and George E. Fox were two of the people who pioneered the use of 16S rRNA in phylogenetics in 1977.[3] Multiple sequences of the 16S rRNA gene can exist within a single bacterium.

Functions

- Like the large (23S) ribosomal RNA, it has a structural role, acting as a scaffold defining the positions of the ribosomal proteins.
- The 3′-end contains the anti-Shine-Dalgarno sequence, which binds upstream to the AUG start codon on the mRNA. The 3′-end of 16S RNA binds to the proteins S1 and S21 which are known to be involved in initiation of protein synthesis[5]
- Interacts with 23S, aiding in the binding of the two ribosomal subunits (50S and 30S)
- Stabilizes correct codon-anticodon pairing in the A-site by forming a hydrogen bond between the N1 atom of adenine residues 1492 and 1493 and the 2′OH group of the mRNA backbone.



Universal Primers

The 16S rRNA gene is used for phylogenetic studies[7] as it is highly conserved between different species of bacteria and archaea.[8] Carl Woese (1977) pioneered this use of 16S rRNA.[2] It is suggested that 16S rRNA gene can be used as a reliable molecular clock because 16S rRNA sequences from distantly related bacterial lineages are shown to have similar functionalities.[9] Some thermophilic archaea (e.g. order Thermoproteales) contain 16S rRNA gene introns that are located in highly conserved regions and can impact the annealing of "universal" primers.[10] Mitochondrial and chloroplastic rRNA are also amplified.

The most common primer pair was devised by Weisburg *et al.* (1991)[7] and is currently referred to as 27F and 1492R; however, for some applications shorter amplicons may be necessary, for example for 454 sequencing with titanium chemistry the primer pair 27F-534R covering V1 to V3.[11] Often 8F is used rather than 27F. The two primers are almost identical, but 27F has an M instead of a C. AGAGTTTGATC**M**TGGCTCAG compared with 8F

The types of 16S rRNA sequence analysis

The analysis method of 16S rRNA gene fragment mainly includes the following 3 kinds:

(1) Sequencing the PCR products on the plasmid vector and comparing with the sequence in the 16S rRNA database to determine its position in the evolutionary tree and identify the possible

species of microorganism in the sample. The information obtained by this method is the most comprehensive, but in the sample Complex sequencing requires extensive sequencing.

(2) Hybridize the PCR products with 16S rRNA specific probes to obtain microbial composition information. In addition, the probe can be directly detected by in situ hybridization with the sample. In situ hybridization can not only determine the morphological characteristics and abundance of microbes, but also analyze their spatial distribution.

(3) The restriction fragment length polymorphism analysis of PCR products was carried out. The ribose type of microorganism gene was determined by observing the enzyme cut electrophoresis atlas and numerical analysis, and then compared with the data in the ribosome library, and the relationship between the microbial composition of the samples and the species of different microorganisms was analyzed.

Application of 16S rRNA

16S rRNA is the most conservative gene for all bacteria, and some of the most conservative genes in the evolutionary process. Some of the gene sequences remain stable in the long course of evolution. In addition, based on the multiple copies of the bacterial chromosomes, based on the 16S rRNA gene, the PCR, nested PCR, multiple semi nested PCR, RT-PCR, and oligosaccharides are established. Nucleotide probes have been applied to the identification of clinical bacteria, sequence analysis, molecular classification of bacteria, and phylogenetic analysis.

The future of 16S rRNA

Ribosomal rRNA is essential for the survival of all living things. 16 S rRNA is highly conserved in the evolutionary process of bacteria and other microorganisms. It is called “the molecular fossil” of bacteria. At the same time, its conservatism is relative. There are 9 to 10 variation regions (V1 ~ V10) between the conservative areas. There are different degrees of difference in the families, genera and species of different bacteria, so 16S rRNA can be used as both It is a marker for bacterial classification and can be used as a target molecule for detection and identification of clinical pathogens. The PCR of the bacterial ribosome 16S rRNA gene as the target molecule can judge the existence of bacterial infection early and identify the species of the pathogen by further analysis of the amplified products and make up for the above deficiencies. It is an important breach in the diagnosis of infectious diseases and has become the principal of bacteriologists at home and abroad. One of the directions is to be studied.

18srRNA

18S ribosomal RNA (abbreviated **18S rRNA**) is a part of the ribosomal RNA. The S in 18S represents Svedberg units. 18S rRNA is an SSU rRNA, a component of the eukaryotic ribosomal small subunit (40S). 18S rRNA is the structural RNA for the small component of eukaryotic cytoplasmic ribosomes, and thus one of the basic components of all eukaryotic cells.

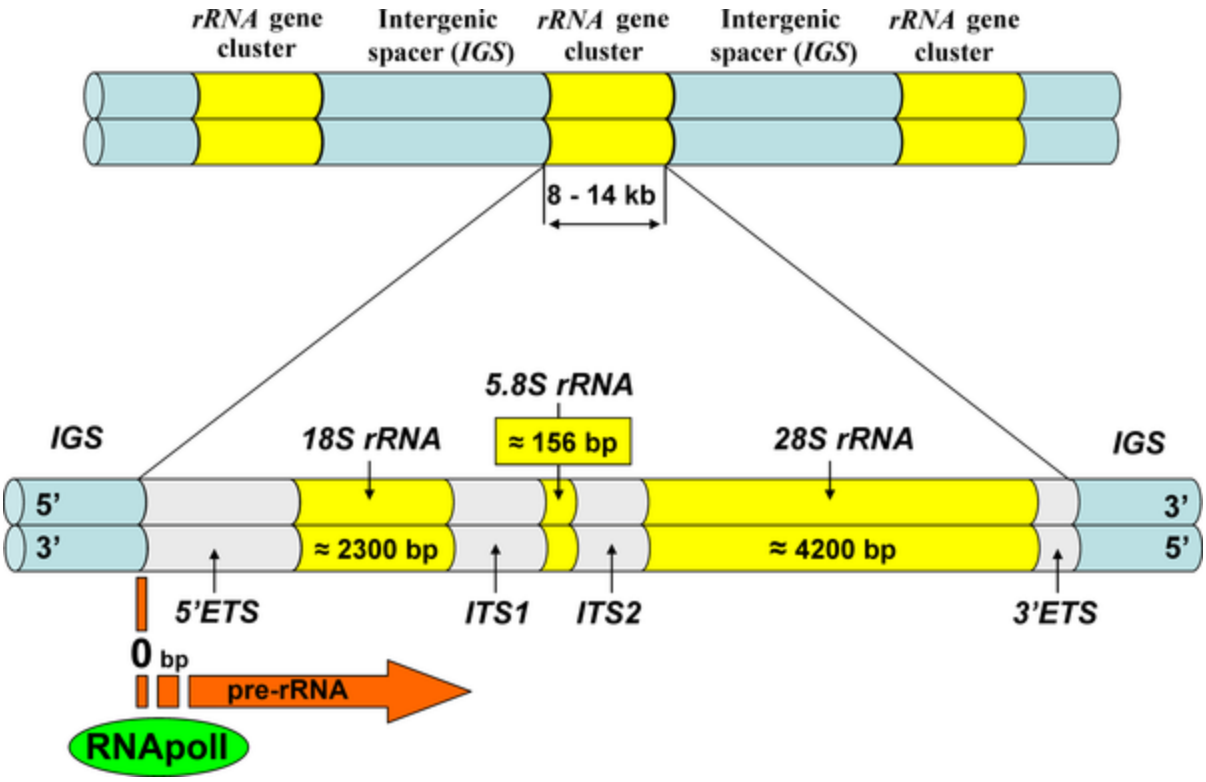
18S rRNA is the eukaryotic cytosolic homologue of 16S ribosomal RNA in prokaryotes and mitochondria.

The genes coding for 18S rRNA are referred to as **18S rRNA genes**. Sequence data from these genes is widely used in molecular analysis to reconstruct the evolutionary history of organisms, especially in vertebrates, as its slow evolutionary rate makes it suitable to reconstruct ancient divergences.

Uses

The small subunit (SSU) 18S rRNA gene is one of the most frequently used genes in phylogenetic studies and an important marker for random target polymerase chain reaction (PCR) in environmental biodiversity screening.[1] In general, rRNA gene sequences are easy to access due to highly conserved flanking regions allowing for the use of universal primers.[1] Their repetitive arrangement within the genome provides excessive amounts of template DNA for PCR, even in the smallest organisms. The 18S gene is part of the ribosomal functional core and is exposed to similar selective forces in all living beings. Thus, when the first large-scale phylogenetic studies based on 18S sequences were published (*e.g.* by Field *et al.*, 1988),[2] the gene was celebrated as the prime candidate for reconstructing the metazoan tree of life.[1] 18S sequences later provided evidence for the splitting of Ecdysozoa and Lophotrochozoa clades (monophyletic group of organisms composed of a common ancestor and all its lineal descendants), thus contributing to the most recent revolutionary change in our understanding of metazoan relationships.[1]

During the latter part of the 2000s, and with increased numbers of taxa included into molecular phylogenies, however, two problems became apparent. First, there are prevailing sequencing impediments in representatives of certain taxa, such as the mollusk classes Solenogastres and Tryblidia, selected bivalve taxa, and the enigmatic crustacean class Remipedia.[1] Failure to obtain 18S sequences of single taxa is considered a common phenomenon but is rarely ever reported.[1] Secondly, in contrast to initially high hopes, 18S cannot resolve nodes at all taxonomic levels and its efficacy varies considerably among clades. This has been discussed as an effect of rapid ancient radiation within short periods. Multigene analyses are currently thought to give more reliable results for tracing deep branching events in Metazoa but 18S still is extensively used in phylogenetic analyses



ITS rRNA

Internal transcribed spacer (**ITS**) is the spacer DNA situated between the small-subunit ribosomal RNA (rRNA) and large-subunit rRNA genes in the chromosome or the corresponding transcribed region in the polycistronic rRNA precursor transcript.

In bacteria and archaea, there is a single ITS, located between the 16S and 23S rRNA genes. Conversely, there are two ITSs in eukaryotes: **ITS1** is located between 18S and 5.8S rRNA genes, while **ITS2** is between 5.8S and 28S (in opisthokonts, or 25S in plants) rRNA genes. ITS1 corresponds to the ITS in bacteria and archaea, while ITS2 originated as an insertion that interrupted the ancestral 23S rRNA gene

In bacteria and archaea, the ITS occurs in one to several copies, as do the flanking 16S and 23S genes. When there are multiple copies, these do not occur adjacent to one another. Rather, they occur in discrete locations in the circular chromosome. It is not uncommon in bacteria to carry tRNA genes in the ITS.[3][4]

In eukaryotes, genes encoding ribosomal RNA and spacers occur in tandem repeats that are thousands of copies long, each separated by regions of non-transcribed DNA termed *intergenic spacer* (IGS) or *non-transcribed spacer* (NTS).

Each eukaryotic ribosomal cluster contains the 5' external transcribed spacer (5' ETS), the 18S rRNA gene, the ITS1, the 5.8S rRNA gene, the ITS2, the 26S or 28S rRNA gene, and finally the 3' ETS.[5]

During rRNA maturation, ETS and ITS pieces are excised. As non-functional by-products of this maturation, they are rapidly degraded

Uses

Sequence comparison of the eukaryotic ITS regions is widely used in taxonomy and molecular phylogeny because of several favorable properties:[7]

- It is routinely amplified thanks to its small size associated to the availability of highly conserved flanking sequences.
- It is easy to detect even from small quantities of DNA due to the high copy number of the rRNA clusters.
- It undergoes rapid concerted evolution via unequal crossing-over and gene conversion. This promotes intra-genomic homogeneity of the repeat units, although high-throughput sequencing showed the occurrence of frequent variations within plant species.[8]
- It has a high degree of variation even between closely related species. This can be explained by the relatively low evolutionary pressure acting on such non-coding spacer sequences.

Mycological barcoding

The ITS region is the most widely sequenced DNA region in molecular ecology of fungi[28] and has been recommended as the universal fungal barcode sequence.[29] It has typically been most useful for molecular systematics at the species to genus level, and even within species (e.g., to identify geographic races). Because of its higher degree of variation than other genic regions of rDNA (for example, small- and large-subunit rRNA), variation among individual rDNA repeats can sometimes be observed within both the ITS and IGS regions. In addition to the universal ITS1+ITS4 primers[30][31] used by many labs, several taxon-specific primers have been described that allow selective amplification of fungal sequences (e.g., see Gardes & Bruns 1993 paper describing amplification of basidiomycete ITS sequences from mycorrhiza samples).[32] Despite shotgun sequencing methods becoming increasingly utilized in microbial sequencing, the low biomass of fungi in clinical samples make the ITS region amplification an area of ongoing research

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 BIOINFORMATICS

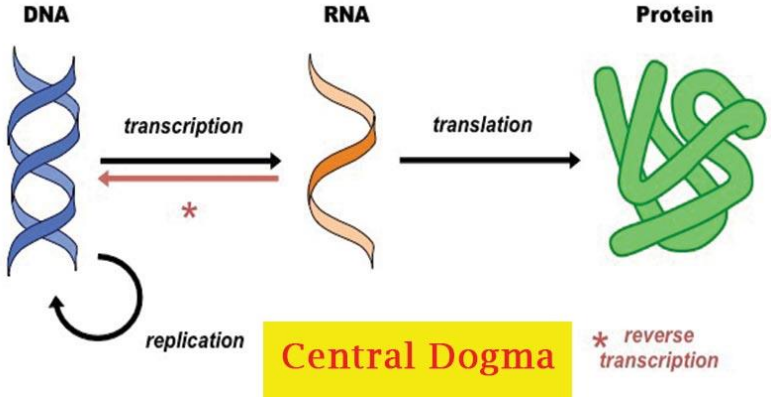
UNIT – 2- SBIA1301 – Molecular Biology and Genomics

Central Dogma

- **DNA** contains the complete genetic information that defines the structure and function of an organism.
- Proteins are formed using the **genetic code** of the DNA.
- Conversion of DNA encoded information to **RNA** is essential to form **proteins**.
- Thus, within most cells, the genetic information flows from – DNA to RNA to protein.
- The flow of information is followed through three different processes which are responsible for the inheritance of genetic information and for its conversion from one form to another:

1. **Replication:** a double stranded nucleic acid is duplicated to give identical copies. This process perpetuates the genetic information.
2. **Transcription:** a DNA segment that constitutes a gene is read and transcribed into a single stranded sequence of RNA. The RNA moves from the nucleus into the cytoplasm.
3. **Translation:** the RNA sequence is translated into a sequence of amino acids as the protein is formed. During translation, the ribosome reads three bases (a codon) at a time from the RNA and translates them into one **amino acid**.

- This flow of information is unidirectional and irreversible.



This explanation is the simplest way in which the **Central Dogma of Molecular Biology** is interpreted.

- In the bigger picture, the central dogma of molecular biology is an explanation of the flow of genetic information within a biological system.
- It was first stated by Francis Crick in 1958, as
- "Once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible."

The Dogmas

- The dogma is a framework for understanding the transfer of sequence information between information-carrying biopolymers, DNA and RNA (both nucleic acids), and protein.
- There are 3×3=9 conceivable direct transfers of information that can occur between these.
- The dogma classes these into 3 groups of 3:

A. Three general transfers

- It describes the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA information can be copied into mRNA (transcription), and proteins can be synthesized using the information in mRNA as a template (translation).
- It is believed to occur normally in most cells.

B. Three special transfers

- The special transfers describe: RNA being copied from RNA (RNA replication), DNA being synthesised using an RNA template (reverse transcription), and proteins being synthesised directly from a DNA template without the use of mRNA.
- Temin (1970) reported the existence of an enzyme "RNA dependent DNA polymerase" (inverse transcriptase) which could synthesize DNA from a single stranded RNA template.
- Baltimore (1970) also reported the activity of this enzyme in certain RNA tumour viruses.
- This exciting finding in molecular biology gave rise to the concept of **central dogma reverse**" or teminism, suggesting that the sequence of information flow is not necessarily from DNA to RNA to protein but can also take place from RNA to DNA.
- It is known to occur, but only under specific conditions in case of some viruses or in a laboratory.

C. Three unknown transfers

- The unknown transfers describe: a protein being copied from a protein, synthesis of RNA using the primary structure of a protein as a template, and DNA synthesis using the primary structure of a protein as a template
- These are not thought to naturally occur.

Significance of the Central Dogma of Molecular Biology

Thus, the central dogma provides the basic framework for how genetic information flows from a DNA sequence to a protein product inside cells and thus give an insight to the important processes going on inside the cells.

DNA replication

UNIT – 3- SBIA1301 – Molecular Biology and Genomics

DNA sequencing

DNA sequencing is the determination of the precise sequence of nucleotides in a sample of DNA. Before the development of direct DNA sequencing methods, DNA sequencing was difficult and indirect. The DNA had to be converted to RNA, and limited RNA sequencing could be done by the existing cumbersome methods. Thus, only shorter DNA sequences could be determined by this method. Using this method, Walter Gilbert and Alan Maxam at Havard University determined that the Lac operator is a 27 bp long sequence.

When a particular gene of interest or a DNA fragment is isolated, the sequencing of that piece of DNA becomes essential. The sequence of DNA refers to the order of nucleotide bases along its sugar phosphate backbone.

No technique can determine the sequence of bases in an entire gene in a single experiment, so it is necessary to cut the whole gene into fragments of manageable size (few hundred base pair long) and purify each fragment. There are two different methods which are now routinely used for determination of DNA sequences.

DNA sequencing is the process of determining the nucleic acid sequence – the order of nucleotides in DNA. It includes any method or technology that is used to determine the order of the four bases: adenine, guanine, cytosine, and thymine. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.[1][2]

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. Comparing healthy and mutated DNA sequences can diagnose different diseases including various cancers,[3] characterize antibody repertoire,[4] and can be used to guide patient treatment.[5] Having a quick way to sequence DNA allows for faster and more individualized medical care to be administered, and for more organisms to be identified and cataloged.[4]

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.

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,[6] DNA sequencing has become easier and orders of magnitude faster

Two main methods are widely known to be used to sequence DNA:

1. **The Chemical Method** (also called the Maxam–Gilbert method after its inventors).
2. **The Chain Termination Method** (also known as the Sanger dideoxy method after its inventor).

- Maxam–Gilbert technique depends on the relative chemical liability of different nucleotide bonds, whereas the Sanger method interrupts elongation of DNA sequences by incorporating dideoxynucleotides into the sequences.
- The chain termination method is the method more usually used because of its speed and simplicity.

Maxam Gilbert method / Chemical degradation method

- In 1976-1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases.
- The method requires radioactive labelling at one end and purification of the DNA fragment to be sequenced.
- Chemical treatment generates breaks at a small proportions of one or two of the four nucleotide based in each of four reactions (G,A+G, C, C+T).
- Thus a series of labelled fragments is generated, from the radiolabelled end to the first 'cut' site in each molecule.
- The fragments in the four reactions are arranged side by side in gel electrophoresis for size separation.
- To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabelled DNA fragment, from which the sequence may be inferred.

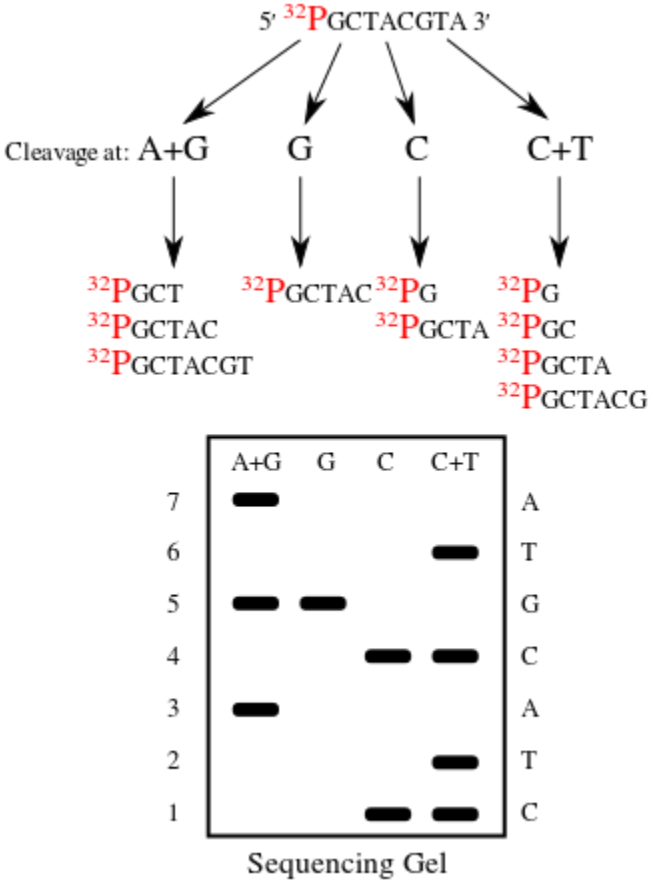
Key features

- Base-specific cleavage of DNA by certain chemicals
- Four different chemicals, one for each base
- A set of DNA fragments of different sizes
- DNA fragments contain up to 500 nucleotides

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-^{32}P 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; $(CH_2)_5NH$ 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



Advantages

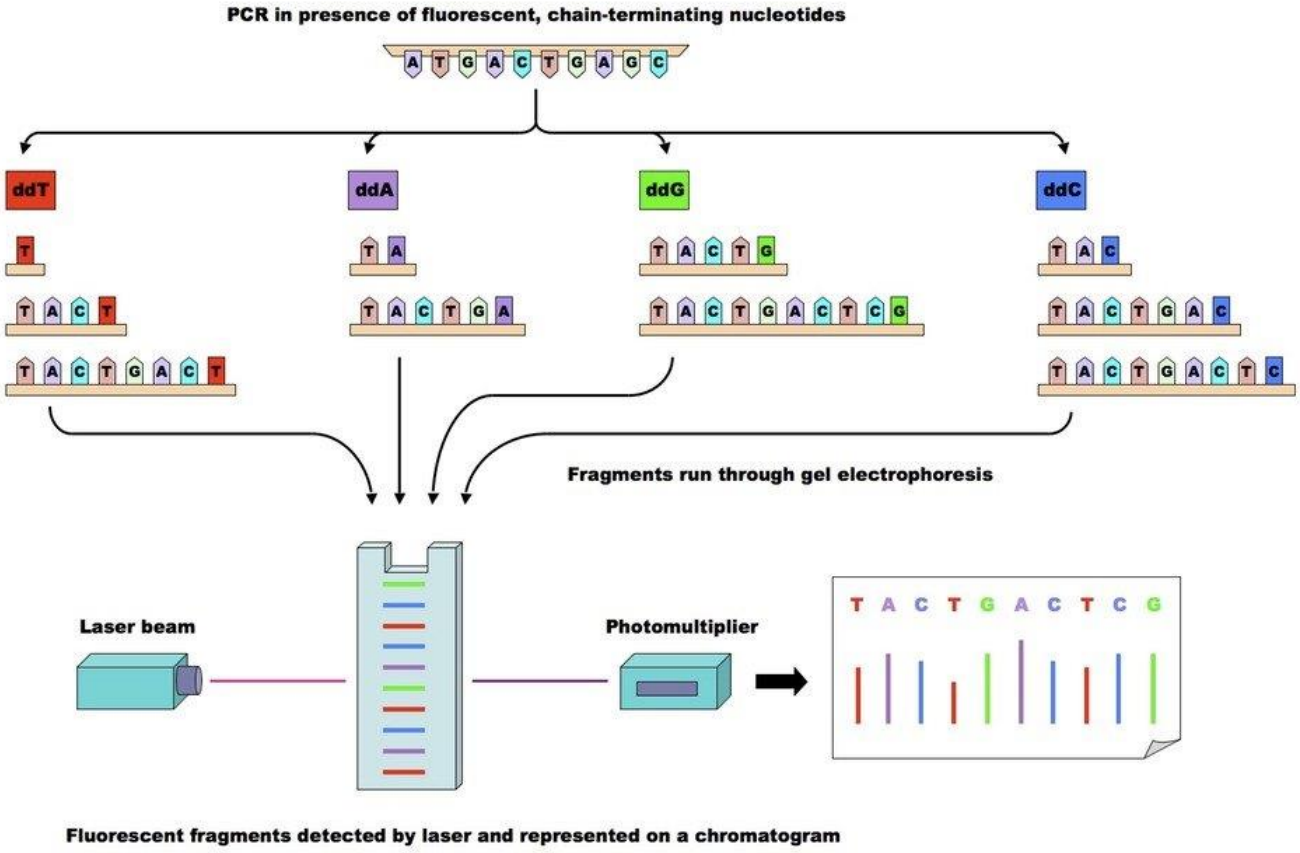
- Purified DNA can be read directly
- Homopolymeric DNA runs are sequenced as efficiently as heterogeneous DNA sequences
- Can be used to analyze DNA protein interactions (i.e. footprinting)
- Can be used to analyze nucleic acid structure and epigenetic modifications to DNA

Disadvantages

- It requires extensive use of hazardous chemicals.
- It has a relatively complex set up / technical complexity.
- It is difficult to “scale up” and cannot be used to analyze more than 500 base pairs.
- The read length decreases from incomplete cleavage reactions.
- It is difficult to make Maxam-Gilbert sequencing based DNA kits.

Sangers method / Chain termination method

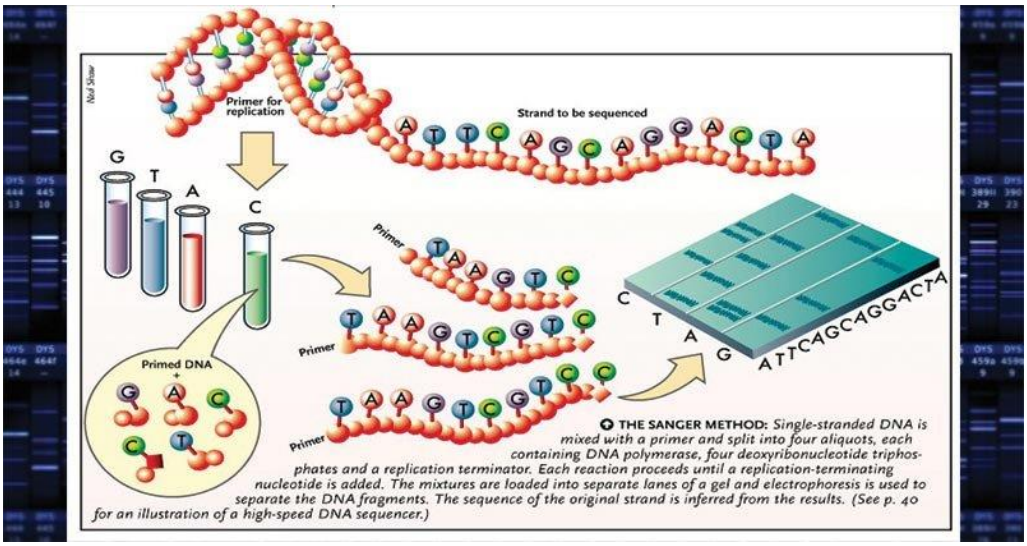
- Sanger's method of gene sequencing is also known as dideoxy chain termination method. It generates nested set of labelled fragments from a template strand of DNA to be sequenced by replicating that template strand and interrupting the replication process at one of the four bases.
- Four different reaction mixtures are produced that terminates in A. T. G or C respectively.



- The chain terminator method is more efficient and uses fewer toxic chemicals and lower amount of radioactivity than the method of Maxam and Gilbert.
- The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.
- The chain termination method requires a single-stranded DNA template, a DNA primer,a DNA polymerase, radioactively or fluorescently labelled nucleotides,and modified nucleotides that terminate DNA strand elongation.
- The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP, dTTP) and the DNA polymerase.
- To each reaction is added only one of the four dideoxynucleotide (ddATP, ddGTP, ddCTP, ddTTP) which are the chain terminating nucleotides, lacking a 3'-OH group required for the

formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

- The newly synthesized and labelled DNA fragments are heat denatured, and separated by size by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of the four individual lanes (lanes A, T, G, C).
- The DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image.
- A dark band in a lane indicates a DNA fragment that is result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP).
- The relative position of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.
- The technical variations of chain termination sequencing include tagging with nucleotides containing radioactive phosphorus for labelling, or using a primer labelled 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.



Key Features

- Uses dideoxy nucleotides to terminate DNA synthesis.
- DNA synthesis reactions in four separate tubes
- Radioactive dATP is also included in all the tubes so the DNA products will be radioactive.
- Yielding a series of DNA fragments whose sizes can be measured by electrophoresis.
- The last base in each of these fragments is known.

Advantage

Chain termination methods have greatly simplified DNA sequencing.

Limitations

- Non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence.
- DNA secondary structures affecting the fidelity of the sequence.

Significance of DNA sequencing

- Information obtained by DNA sequencing makes it possible to understand or alter the function of genes.
- DNA sequence analysis demonstrates regulatory regions that control gene expression and genetic "hot spots" particularly susceptible to mutation.
- Comparison of DNA sequences shows evolutionary relationships that provide a framework for definite classification of microorganisms including viruses.
- Comparison of DNA sequences facilitates identification of conserved regions, which are useful for development of specific hybridization probes to detect microorganisms including viruses in clinical samples.
- DNA sequencing has become sufficiently fast and inexpensive to allow laboratory determination of microbial sequences for identification of microbes. Sequencing of the 16S ribosomal subunit can be used to identify specific bacteria. Sequencing of viruses can be used to identify the virus and distinguish different strains.
- DNA sequencing shows gene structure that helps research workers to find out the structure of gene products.

Automated DNA sequencing

Large-scale DNA sequencing requires automated procedures based on fluorescence labeling of DNA and suitable detection systems. In general, a fluorescent label can be used either directly or indirectly. Direct fluorescent labels, as used in automated sequencing, are fluorophores. These are molecules that emit a distinct fluorescent color when exposed to UV light of a specific wavelength. Examples of fluorophores used in sequencing are fluorescein, which fluoresces pale green when exposed to a wavelength of 494 nm; rhodamine, which fluoresces red at 555 nm; and aminomethyl cumarin acetic acid, which fluoresces blue at 399 nm. In addition, a combination of different fluorophores can be used to produce a fourth color. Thus, each of the four bases can be distinctly labeled.

Another approach is to use PCR-amplified products (thermal cycle sequencing). This has the advantage that double-stranded rather than single-stranded DNA can be used as the starting material. And since small amounts of template DNA are sufficient, the DNA to be sequenced does not have to be cloned beforehand.

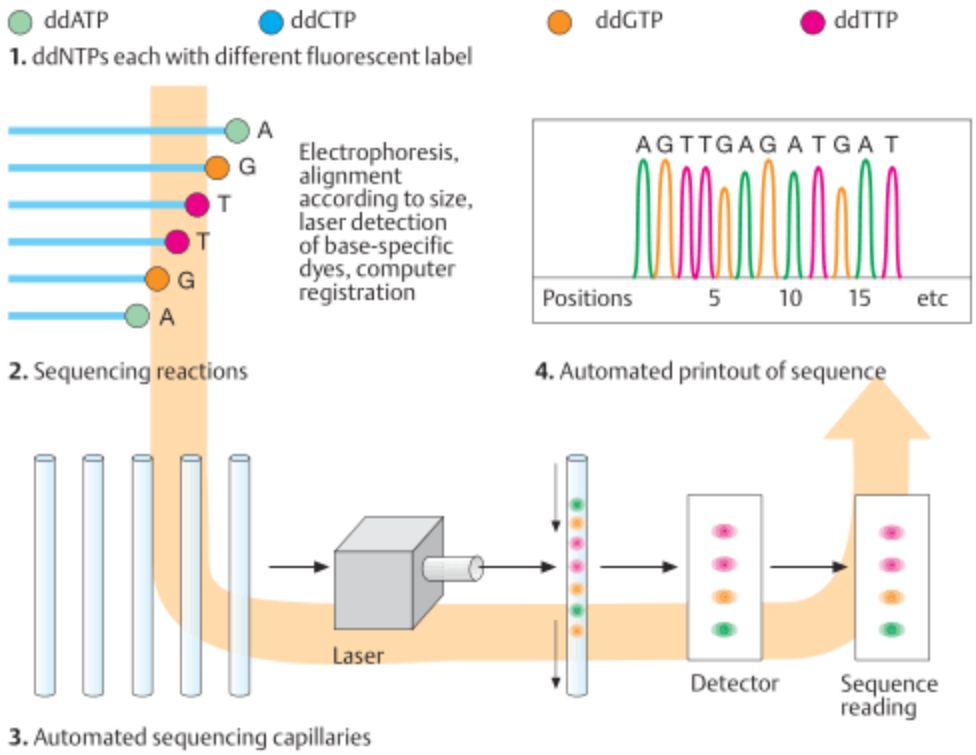
Thermal cycle sequencing

The DNA to be sequenced is contained in vector DNA [1]. The primer, a short oligonucleotide with a sequence complementary to the site of attachment on the single-stranded DNA, is used as a starting point. For sequencing short stretches of DNA, a universal primer is sufficient.This is an oligonucleotide that will bind to vector DNA adjacent to the DNA to be sequenced. However, if the latter is longer than about 750 bp, only part of it will be sequenced. Therefore, additional internal primers are required. These anneal to different sites and amplify the DNA in a series of contiguous, overlapping chain termination experiments

Here, each primer determines which region of the template DNA is being sequenced. In thermal cycle sequencing , only one primer is used to carry out PCR reactions, each with one dideoxynucleotide (ddA, ddT, ddG, or ddC) in the reaction mixture. This generates a series of different chain-terminated strands, each dependent on the position of the particular nucleotide base where the chain is being terminated [4]. After many cycles and with electrophoresis, the sequence can be read as shown in the previous plate. One advantage of thermal cycle sequencing is that double-stranded DNA can be used as starting material.

Automated DNA sequencing (principle)

Automated DNA sequencing involves four fluorophores, one for each of the four nucleotide bases. The resulting fluorescent signal is recorded at a fixed point when DNA passes through a capillary containing an electrophoretic gel. The base-specific fluorescent labels are attached to appropriate dideoxynucleotide triphosphates (ddNTP). Each ddNTP is labeled with a different color, e.g., ddATP green, ddCTP blue, ddGTP yellow, and ddTTP red [5]. (The actual colors for each nucleotide may be different.) All chains terminated at an adenine (A) will yield a green signal; all chains terminated at a cytosine (C) will yield a blue signal, and so on. The sequencing reactions based on this kind of chain termination at labeled nucleotides [6] are carried out automatically in sequencing capillaries [7]. The electrophoretic migration of the ddNTP-labeled chains in the gel in the capillary pass in front of a laser beam focused on a fixed position. The laser induces a fluorescent signal that is dependent on the specific label representing one of the four nucleotides. The sequence is electronically read and recorded and is visualized as alternating peaks in one of the four colors, representing the alternating nucleotides in their sequence positions. In practice the peaks do not necessarily show the same maximal intensity as in the schematic diagram shown here. (Illustration based on Brown, 1999, and Strachan and Read, 1999).



Pyro sequencing

Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA) based on the "sequencing by synthesis" principle, in which the sequencing is performed by detecting the nucleotide incorporated by a DNA polymerase. Pyrosequencing relies on light detection based on a chain reaction when pyrophosphate is released. Hence, the name pyrosequencing.

The principle of Pyrosequencing was first described in 1993[1] by Bertil Pettersson, Mathias Uhlen and Pål Nyren by combining the solid phase sequencing method using streptavidin coated magnetic beads with recombinant DNA polymerase lacking 3´to 5´exonuclease activity (proof-reading) and luminescence detection using the firefly luciferase enzyme.[3] A mixture of three enzymes (DNA polymerase, ATP sulfurylase and firefly luciferase) and a nucleotide (dNTP) are added to single stranded DNA to be sequenced and the incorporation of nucleotide is followed by measuring the light emitted. The intensity of the light determines if 0, 1 or more nucleotides have been incorporated, thus showing how many complementary nucleotides are present on the template strand. The nucleotide mixture is removed before the next nucleotide mixture is added. This process is repeated with each of the four nucleotides until the DNA sequence of the single stranded template is determined.

A second solution-based method for Pyrosequencing was described in 1998[4] by Mostafa Ronaghi, Mathias Uhlen and Pål Nyren. In this alternative method, an additional enzyme apyrase is introduced to remove nucleotides that are not incorporated by the DNA

polymerase. This enabled the enzyme mixture including the DNA polymerase, the luciferase and the apyrase to be added at the start and kept throughout the procedure, thus providing a simple set-up suitable for automation. An automated instrument based on this principle was introduced to the market the following year by the company Pyrosequencing.

A third microfluidic variant of the Pyrosequencing method was described in 2005[5] by Jonathan Rothberg and co-workers at the company 454 Life Sciences. This alternative approach for Pyrosequencing was based on the original principle of attaching the DNA to be sequenced to a solid support and they showed that sequencing could be performed in a highly parallel manner using a microfabricated microarray. This allowed for high-throughput DNA sequencing and an automated instrument was introduced to the market. This became the first next generation sequencing instrument starting a new era in genomics research, with rapidly falling prices for DNA sequencing allowing whole genome sequencing at affordable prices.

Procedure

"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 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.[6]

For the solution-based version of Pyrosequencing, 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.

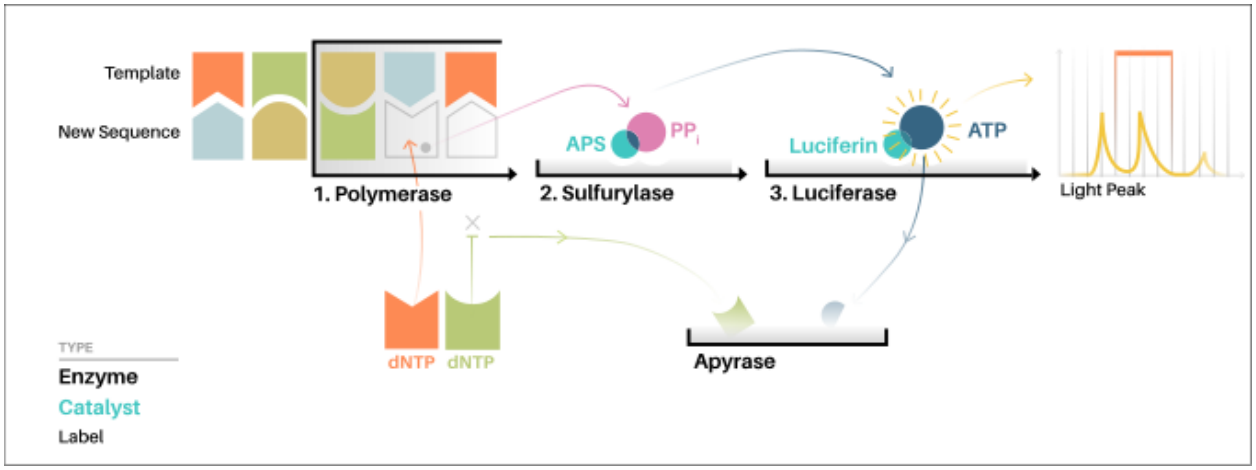
1. The addition of one of the four deoxynucleotide triphosphates (dNTPs) (dATPαS, which is not a substrate for a luciferase, is added instead of dATP to avoid noise) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases 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. The light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a program.
3. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide.

The process can be represented by the following equations:

- PPi + APS → ATP + Sulfate (catalyzed by ATP-sulfurylase);
- ATP + luciferin + O2 → AMP + PPi + oxyluciferin + CO2 + hv (catalyzed by luciferase);

where:

- PPi is pyrophosphate
- APS is adenosine 5-phosphosulfate;
- ATP is adenosine triphosphate;
- O2 is oxygen molecule;
- AMP is adenosine monophosphate;
- CO2 is carbon dioxide;
- hv is light.



Limitations

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. Lack of proof-reading activity limits accuracy of this method.

Shotgun sequencing

In genetics, **shotgun sequencing** is a method used for sequencing random DNA strands. It is named by analogy with the rapidly expanding, quasi-random shot grouping of a shotgun.

The chain-termination method of DNA sequencing ("Sanger sequencing") can only be used for short DNA strands of 100 to 1000 base pairs. Due to this size limit, longer sequences are subdivided into smaller fragments that can be sequenced separately, and these sequences are assembled to give the overall sequence.

There are two principal methods for this fragmentation and sequencing process. Primer walking (or "chromosome walking") progresses through the entire strand piece by piece, whereas shotgun sequencing 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 whole genome sequencing.

Large, mammalian genomes? are particularly difficult to clone?, sequence and assemble because of their size and structural complexity. As a result clone-by-clone sequencing, although reliable and methodical, takes a very long time. With the emergence of cheaper sequencing? and more sophisticated computer programs, researchers have therefore relied on whole genome shotgun sequencing to tackle larger, more complex genomes.

- Shotgun sequencing was originally used by Fred Sanger and his colleagues to sequence small genomes such as those of viruses? and bacteria?.
- Whole genome shotgun sequencing bypasses the time-consuming mapping and cloning steps that make clone-by-clone sequencing so slow.
- In whole genome shotgun sequencing the entire genome is broken up into small fragments of DNA? for sequencing.
- These fragments are often of varying sizes, ranging from 2-20 kilobases? (2,000-20,000 base pairs?) to 200-300 kilobases (200,000-300,000 base pairs).
- These fragments are sequenced to determine the order of the DNA bases?, A, C, G and T.
- The sequenced fragments are then assembled together by computer programs that find where fragments overlap.
- You can imagine shotgun sequencing as being a bit like shredding multiple copies of a book (which in this case is a genome), mixing up all the fragments and then reassembling the original text (genome) by finding fragments with text that overlap and piecing the book back together again.
- This method of genome sequencing was used by Craig Venter, founder of the private company Celera Genomics, to sequence the human genome. Venter wanted to sequence the human genome faster than the publicly funded effort and felt this was the best way. To assemble the sequence Venter used the clone-by-clone publically available data from the Human Genome Project.
- Now, as technologies are improving, whole genome shotgun sequencing is being used to improve the accuracy of existing genome sequences, such as the reference human genome.
- It is used to remove errors, fill in gaps or correct parts of the sequence that were originally assembled incorrectly when clone-by-clone sequencing was used.

- As a consequence the reference human genome is constantly being improved to ensure that the genome sequence is of the highest possible standard.

What are the advantages of shotgun sequencing?

- By removing the mapping stages, whole genome shotgun sequencing is a much faster process than clone-by-clone sequencing.
- Whole genome shotgun sequencing uses a fraction of the DNA that clone-by-clone sequencing needs.
- Whole genome shotgun sequencing is particularly efficient if there is an existing reference sequence. It is much easier to assemble the genome sequence by aligning it to an existing reference genome[?].
- Shotgun sequencing is much faster and less expensive than methods requiring a genetic map.

What are the disadvantages of shotgun sequencing?

- Vast amounts of computing power and sophisticated software are required to assemble shotgun sequences together. To sequence the genome from a mammal (billions of bases long), you need about 60 million individual DNA sequence reads.
- Errors in assembly are more likely to be made because a genetic map is not used. However these errors are generally easier to resolve than in other methods and minimised if a reference genome can be used.
- Whole genome shotgun sequencing can only really be carried out if a reference genome is already available, otherwise assembly is very difficult without an existing genome to match it to.
- Whole genome shotgun sequencing can also lead to errors which need to be resolved by other, more labour-intensive types of sequencing, such as clone-by-clone sequencing.
- Repetitive genomes and sequences can be more difficult to assemble.

Chromosome walking and jumping

Chromosome walking and chromosome jumping are two technical tools used in molecular biology for locating genes on the chromosomes and physical mapping of the genomes. Chromosome walking is a technique used to clone a target gene in a genomic library by repeated isolation and cloning of adjacent clones of the genomic library. Chromosomal jumping is a special version of chromosomal walking which overcomes the breakpoints of chromosomal walking. Chromosomal walking can only sequence and map small lengths of chromosomes while chromosomal jumping enables sequencing of large parts of chromosomes. This is the key difference between Chromosomal walking and chromosomal jumping.

Chromosome walking

Chromosome walking is a tool which explores the unknown sequence regions of chromosomes by using overlapping restriction fragments. In chromosome walking, a part of a known gene is used as a probe and continued with characterizing the full length of the chromosome to be mapped or sequenced. This goes from the marker to the target length. In chromosome walking, the ends of each overlapping fragments are used for hybridization to identify the next sequence. The probes are prepared from the end pieces of cloned DNA and they are subcloned. Then they are used to find the next overlapping fragment. All these overlapping sequences are used to construct the genetic map of the chromosome and locate the target genes. It is a method of analyzing long stretches of DNA by small overlapping fragments from the recontructed genomic library.

Steps

1. Isolation of a DNA fragment which contains the known gene or marker near target gene
2. Preparation of the restriction map of the selected fragment and subcloning the end region of the fragment to use as a probe
3. Hybridization of the probe with the next overlapping fragment
4. Preparation of the restriction map of the fragment 1 and subcloning of the end region of the fragment 1 to use as a probe for the identification of the next overlapping fragment.
5. Hybridization of the probe with the next overlapping fragment 2
6. Preparation of the restriction map of fragment 2 and subcloning of the end region of the fragment 2 to serve as a probe for the identification of the next overlapping fragment

   Above steps should be continued till the target gene or up to 3’ end of the total length of the sequence.

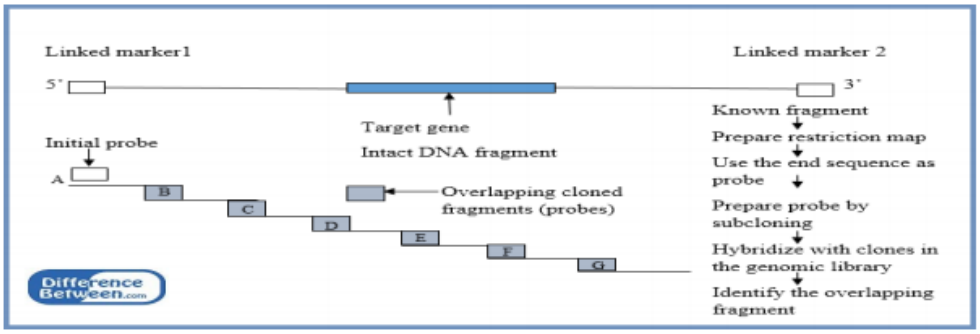


Figure 01: Chromosome Walking Technique

Chromosome walking is an important aspect of cytogenetic in finding SNPs of many organisms and analyzing the genetically transmitted diseases and finding mutations of relevant genes.

Chromosome jumping

Chromosomal jumping is a technique used in molecular biology for physical mapping of genomes of the organisms. This technique was introduced to overcome a barrier of the chromosomal walking which arose upon finding the repetitive DNA regions during the cloning process. Therefore, chromosome jumping technique can be considered as a special version of chromosomal walking. It is a rapid method compared to chromosomal walking and enables bypassing of the repetitive DNA sequences which are not prone to be cloned during chromosomal walking.

Chromosomal jumping narrows the gap between the target gene and the available known markers for genome mapping. Chromosome jumping tool starts with the cutting of a specific DNA with special restriction endonucleases and ligation of the fragments into circularized loops. Then a primer designed from a known sequence is used to sequence the circularized loops. This primer enables jumping and sequencing in an alternative manner. Hence, it can bypass the repetitive DNA sequences and rapidly walk through the chromosome for the search of the target gene.

The discovery of the gene encodes for cystic fibrosis disease was done using the chromosomal jumping tool. Combined together, chromosomal jumping and walking can enhance the genome mapping process.

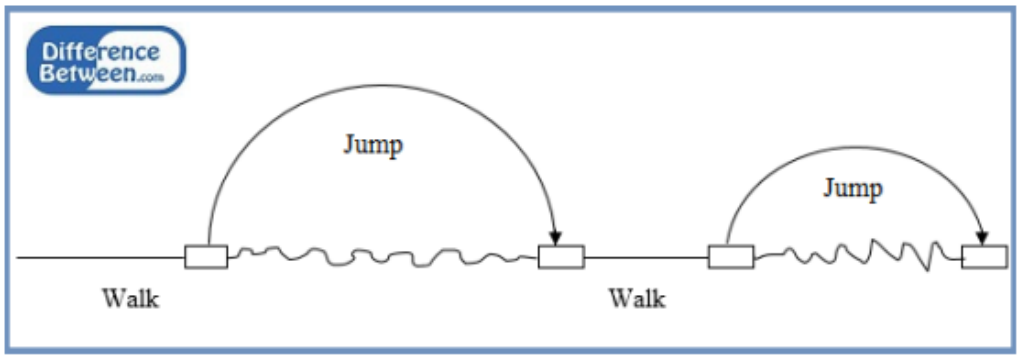


Figure 02: Chromosome Jumping

Summary – Chromosome Walking vs Jumping

Chromosomal walking is frequently applied when it is known that a particular gene is located near a previously cloned gene in a chromosome and it is possible to identify it with repeated isolation of adjacent genomic clones from the genomic library. However, when repetitive DNA regions are found during the chromosomal walking technique, the process cannot be continued. Hence, the technique breaks from that point. Chromosomal jumping is a molecular biological tool which overcomes this limitation for mapping genomes. It bypasses these repetitive DNA regions which are difficult to clone and helps in physical mapping of genomes. This is the main difference between chromosome walking and jumping.

DNA foot printing

DNA footprinting is a method of investigating the sequence specificity of DNA-binding proteins in vitro. This technique can be used to study protein-DNA interactions both outside and within cells.

The regulation of transcription has been studied extensively, and yet there is still much that is not known. Transcription factors and associated proteins that bind promoters, enhancers, or silencers to drive or repress transcription are fundamental to understanding the unique regulation of individual genes within the genome. Techniques like DNA footprinting help elucidate which proteins bind to these associated regions of DNA and unravel the complexities of transcriptional control.

In 1978, David Galas and Albert Schmitz developed the DNA footprinting technique to study the binding specificity of the lac repressor protein. It was originally a modification of the Maxam-Gilbert chemical sequencing technique

Method

The simplest application of this technique is to assess whether a given protein binds to a region of interest within a DNA molecule.[2] Polymerase chain reaction (PCR) amplify and label region of interest that contains a potential protein-binding site, ideally amplicon is between 50 and 200 base pairs in length. Add protein of interest to a portion of the labeled template DNA; a portion should remain separate without protein, for later comparison. Add a cleavage agent to both portions of DNA template. The cleavage agent is a chemical or enzyme that will cut at random locations in a sequence independent manner. The reaction should occur just long enough to cut each DNA molecule in only one location. A protein that specifically binds a region within the DNA template will protect the DNA it is bound to from the cleavage agent. Run both samples side by side on a polyacrylamide gel electrophoresis. The portion of DNA template without protein will be cut at random locations, and thus when it is run on a gel, will produce a ladder-like distribution. The DNA template with the protein will result in ladder distribution with a break in it, the "footprint", where the DNA has been protected from the cleavage agent. Note: Maxam-Gilbert chemical DNA sequencing can be run alongside the samples on the polyacrylamide gel to allow the prediction of the exact location of ligand binding site.

Labeling

The DNA template labeled at the 3' or 5' end, depending on the location of the binding site(s). Labels that can be used are: radioactivity and fluorescence. Radioactivity has been traditionally used to label DNA fragments for footprinting analysis, as the method was originally developed from the Maxam-Gilbert chemical sequencing technique. Radioactive labeling is very sensitive and is optimal for visualizing small amounts of DNA. Fluorescence is a desirable advancement due to the hazards of using radio-chemicals. However, it has been more difficult to optimize because it is not always sensitive enough to detect the low concentrations of the target DNA strands used in DNA footprinting experiments. Electrophoretic sequencing gels or capillary electrophoresis have been successful in analyzing footprinting of fluorescent tagged fragments

Cleaving agent

A variety of cleavage agents can be chosen. a desirable agent is one that is sequence neutral, easy to use, and is easy to control. Unfortunately no available agents meet all of these standards, so an appropriate agent can be chosen, depending on your DNA sequence and ligand of interest. The following cleavage agents are described in detail: DNase I is a large protein that functions as a double-strand endonuclease. It binds the minor groove of DNA and cleaves the phosphodiester backbone. It is a good cleavage agent for footprinting because its size makes it easily physically hindered. Thus is more likely to have its action blocked by a bound protein on a DNA sequence. In addition, the DNase I enzyme is easily controlled by adding EDTA to stop the reaction. There are however some limitations in using DNase I. The enzyme does not cut DNA randomly; its activity is affected by local DNA structure and sequence and therefore results in an uneven ladder. This can limit the precision of predicting a protein’s binding site on the DNA

molecule.[2][3] Hydroxyl radicals are created from the Fenton reaction, which involves reducing Fe^{2+} with H_2O_2 to form free hydroxyl molecules. These hydroxyl molecules react with the DNA backbone, resulting in a break. Due to their small size, the resulting DNA footprint has high resolution. Unlike DNase I they have no sequence dependence and result in a much more evenly distributed ladder. The negative aspect of using hydroxyl radicals is that they are more time consuming to use, due to a slower reaction and digestion time.[4] Ultraviolet irradiation can be used to excite nucleic acids and create photoreactions, which results in damaged bases in the DNA strand.[5] Photoreactions can include: single strand breaks, interactions between or within DNA strands, reactions with solvents, or crosslinks with proteins. The workflow for this method has an additional step, once both your protected and unprotected DNA have been treated, there is subsequent primer extension of the cleaved products.[6][7] The extension will terminate upon reaching a damaged base, and thus when the PCR products are run side-by-side on a gel; the protected sample will show an additional band where the DNA was crosslinked with a bound protein. Advantages of using UV are that it reacts very quickly and can therefore capture interactions that are only momentary. Additionally it can be applied to *in vivo* experiments, because UV can penetrate cell membranes. A disadvantage is that the gel can be difficult to interpret, as the bound protein does not protect the DNA, it merely alters the photoreactions in the vicinity

Applications

In vivo footprinting

In vivo footprinting is a technique used to analyze the protein-DNA interactions that are occurring in a cell at a given time point.[9][10] DNase I can be used as a cleavage agent if the cellular membrane has been permeabilized. However the most common cleavage agent used is UV irradiation because it penetrates the cell membrane without disrupting cell state and can thus capture interactions that are sensitive to cellular changes. Once the DNA has been cleaved or damaged by UV, the cells can be lysed and DNA purified for analysis of a region of interest. Ligation-mediated PCR is an alternative method to footprint *in vivo*. Once a cleavage agent has been used on the genomic DNA, resulting in single strand breaks, and the DNA is isolated, a linker is added onto the break points. A region of interest is amplified between the linker and a gene-specific primer, and when run on a polyacrylamide gel, will have a footprint where a protein was bound.[11] *In vivo* footprinting combined with immunoprecipitation can be used to assess protein specificity at many locations throughout the genome. The DNA bound to a protein of interest can be immunoprecipitated with an antibody to that protein, and then specific region binding can be assessed using the DNA footprinting technique.

Quantitative footprinting

The DNA footprinting technique can be modified to assess the binding strength of a protein to a region of DNA. Using varying concentrations of the protein for the footprinting experiment, the

appearance of the footprint can be observed as the concentrations increase and the proteins binding affinity can then be estimated.[2]

Detection by capillary electrophoresis

To adapt the footprinting technique to updated detection methods, the labelled DNA fragments are detected by a capillary electrophoresis device instead of being run on a polyacrylamide gel. If the DNA fragment to be analyzed is produced by polymerase chain reaction (PCR), it is straightforward to couple a fluorescent molecule such as carboxyfluorescein (FAM) to the primers. This way, the fragments produced by DNaseI digestion will contain FAM, and will be detectable by the capillary electrophoresis machine. Typically, carboxytetramethyl-rhodamine (ROX)-labelled size standards are also added to the mixture of fragments to be analyzed. Binding sites of transcription factors have been successfully identified this way

UNIT – 4- SBIA1301 – Molecular Biology and Genomics

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.

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?

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 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.

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 *Eco*RI, 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, AAGC**C**TA to AAGC**T**TA, 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 ≥ 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 Tm can be determined. A SNP will result in a lower than expected Tm (Howell et al. 1999).

Because DASH genotyping is measuring a quantifiable change in Tm, 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

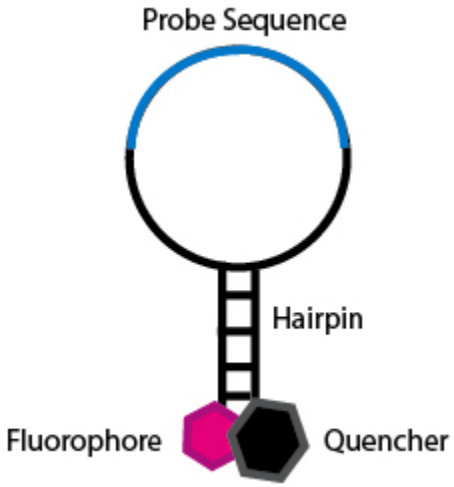
Molecular Beacon Probes

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.

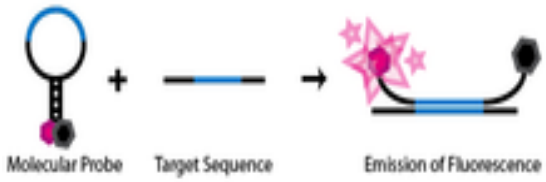
- **5' fluorophore**: Towards the 3' end of the molecular beacon, is attached a dye that fluoresces in presence of a complementary target.
- **3' quencher (non fluorescent)**: The quencher dye is covalently attached to the 3' end of the molecular beacon and when the beacon is in closed loop shape, prevents 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 denatured 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.

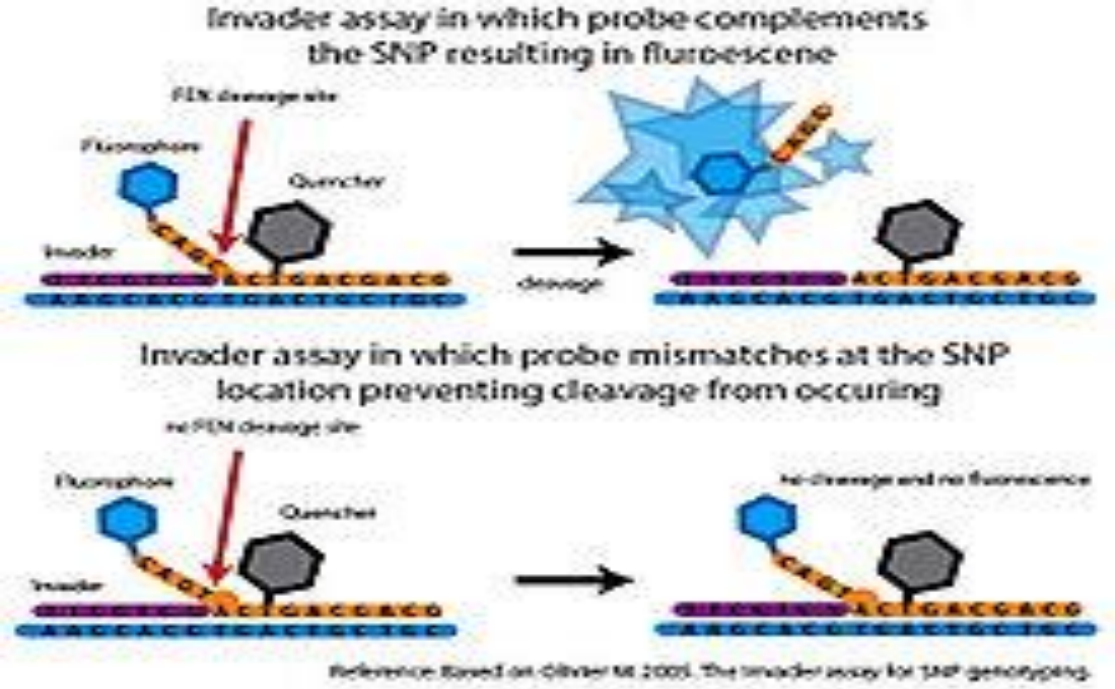
PCR-based methods

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 didioxynucleotide 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 allelle (e.g. from the mother) is missing and instead two copies of the mutant allelle (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 F_1 generation) display the same phenotype. These F_1 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 F_1 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 F_1 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 F_1 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 genotype*AB* 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* × 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 *Eco*RI and *Bam*HI 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 *Eco*RI 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 *Sap*I (5′-GCTCTTC-3′) and *Sgf*I (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 *Bam*HI and *Eco*RI. 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 *Sma*I (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 ^{3}H, 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 A_3, 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

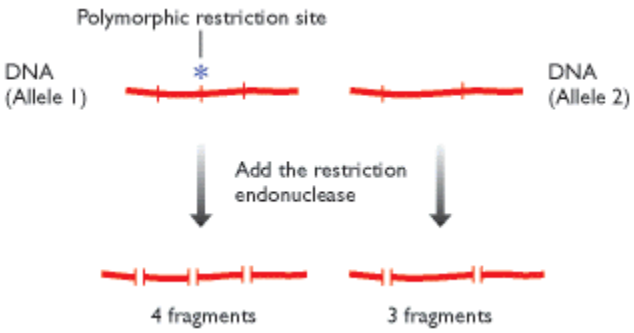


Figure 5.4. A restriction fragment length polymorphism (RFLP). The DNA molecule on the left has a polymorphic restriction site (marked with the asterisk) that is not present in the molecule on the right. The RFLP is revealed after treatment with the restriction enzyme because one of the molecules is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

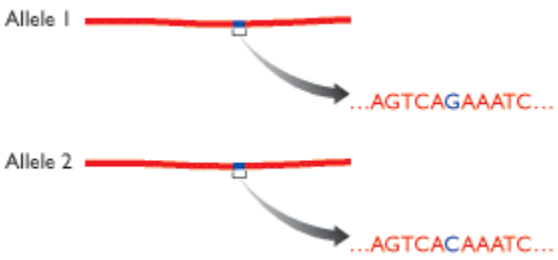


Figure 5.7. A single nucleotide polymorphism (SNP).

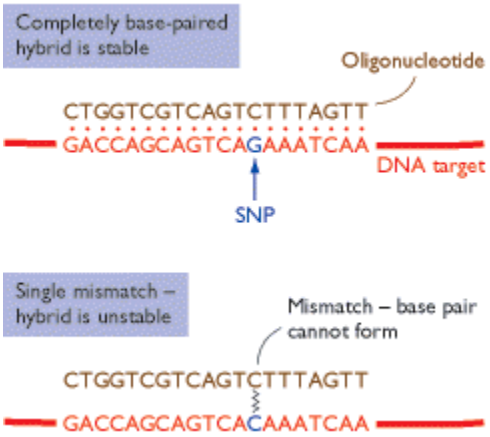


Figure 5.8. Oligonucleotide hybridization is very specific. Under highly stringent hybridization conditions, a stable hybrid occurs only if the oligonucleotide is able to form a completely base-paired structure with the target DNA. If there is a single mismatch then the hybrid does not form. To achieve 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 × number of G and C nucleotides) + (2 × number of A and T nucleotides). This formula gives a rough indication of the T_m for oligonucleotides of 15–30 nucleotides in length.

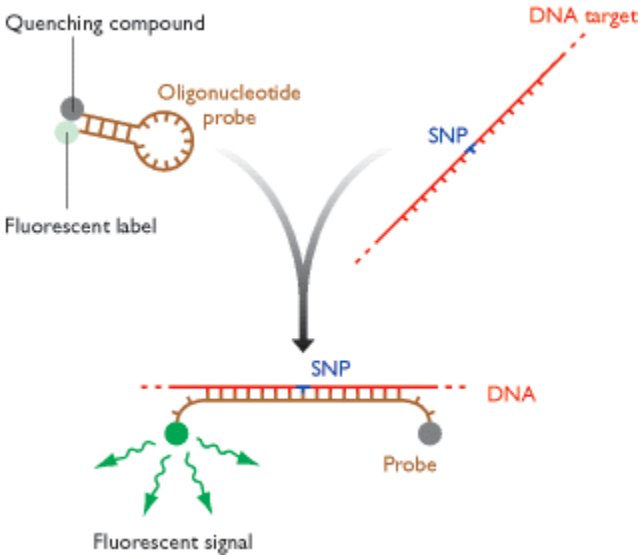


Figure 5.9. One way of detecting an SNP by solution hybridization. The oligonucleotide probe has two end-labels. One of these is a fluorescent dye and the other is a quenching compound. The two ends of the oligonucleotide base-pair to one another, so the fluorescent signal is quenched. When the probe 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'.

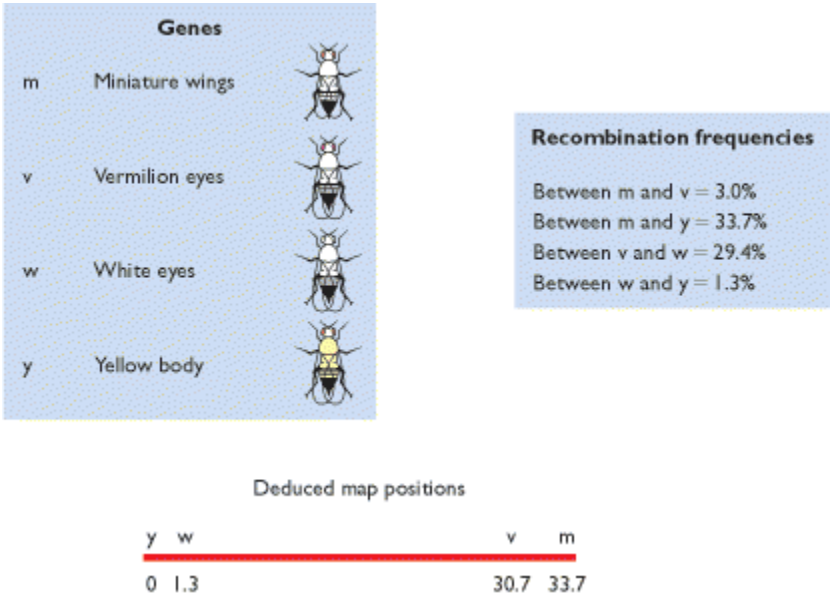


Figure 5.17. Working out a genetic map from recombination frequencies. The example is taken from the original experiments carried out with fruit flies by Arthur Sturtevant. All four genes are

on the X chromosome of the fruit fly. Recombination frequencies between the genes are shown, along with their deduced map positions.

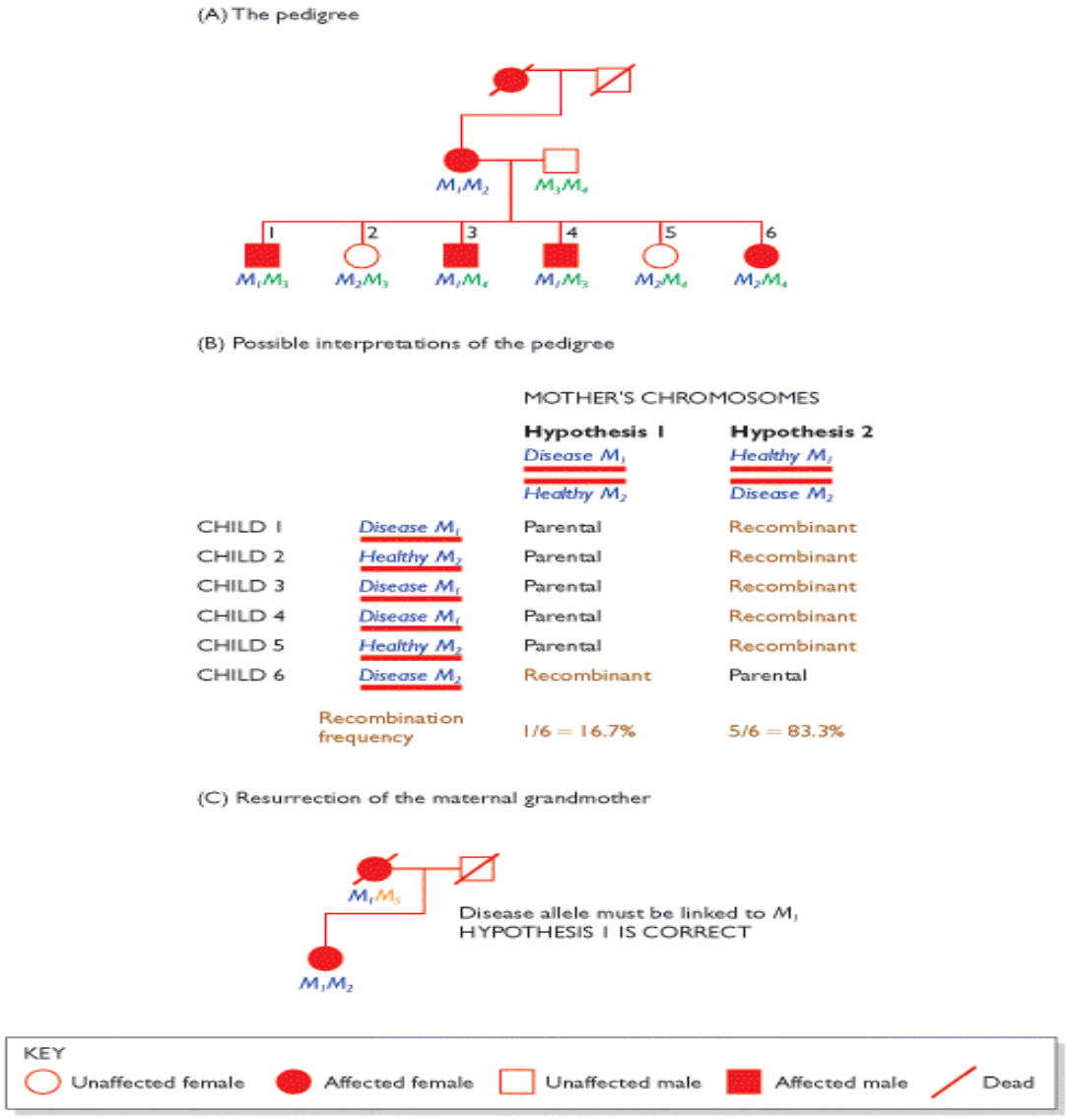
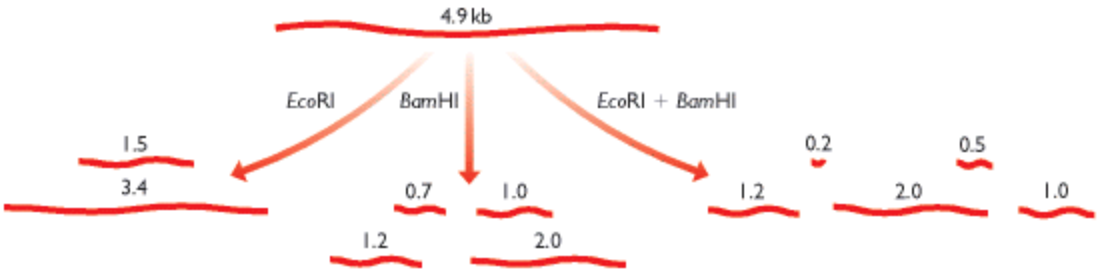


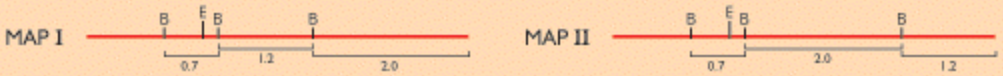
Figure 5.19. An example of human pedigree analysis. (A) The pedigree shows inheritance of a genetic disease in a family of two living parents and six children, with information about the maternal grandparents available from family records. The disease allele (closed symbols) is dominant over the healthy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



INTERPRETATION OF THE DOUBLE RESTRICTION

Fragments	Conclusions
0.2 kb, 0.5 kb	These must derive from the 0.7 kb *Bam*HI fragment, which therefore has an internal *Eco*RI site:
1.0 kb	This must be a *Bam*HI fragment with no internal *Eco*RI site. We can account for the 1.5 kb *Eco*RI fragment if we place the 1.0 kb fragment thus:
1.2 kb, 2.0 kb	These must also be *Bam*HI fragments with no internal *Eco*RI sites. They must lie within the 3.4 kb *Eco*RI fragment. There are two possibilities:



PREDICTED RESULTS OF A PARTIAL *Bam*HI 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.

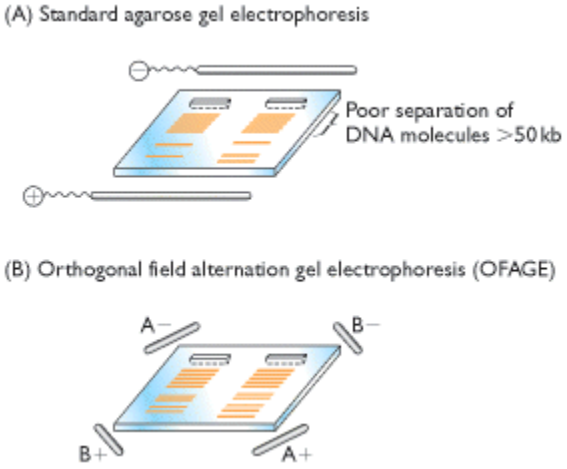


Figure 5.26. Conventional and non-conventional agarose gel electrophoresis. (A) In standard agarose gel electrophoresis the electrodes are placed at either end of the gel and the DNA molecules migrate directly towards the positive electrode. Molecules longer than about 50 kb cannot be separated from 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.

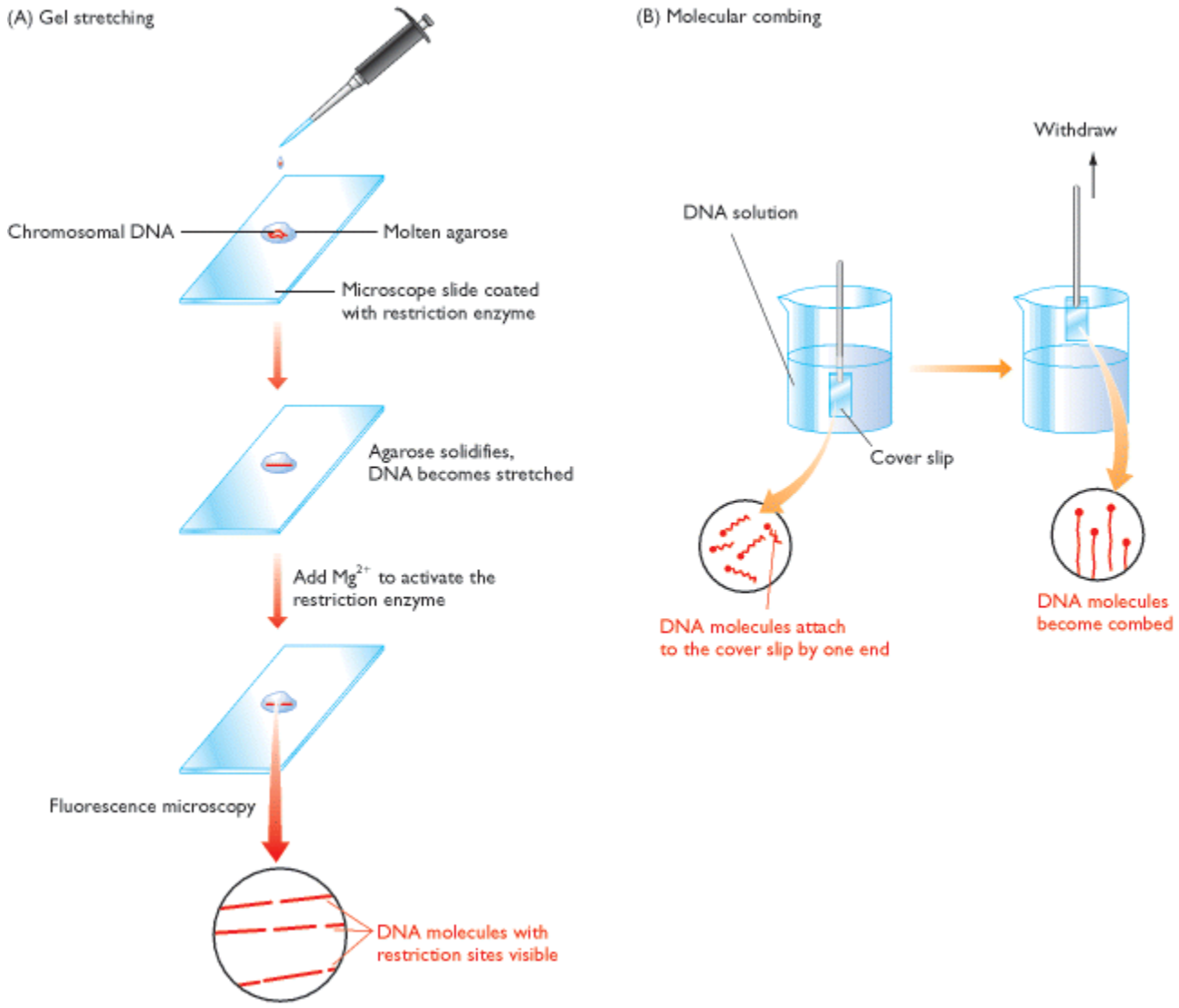


Figure 5.28. Gel stretching and molecular combing. (A) To carry out gel stretching, molten agarose containing chromosomal DNA molecules is pipetted onto a microscope slide coated with a restriction enzyme. As the gel solidifies, the DNA molecules become stretched. It is not understood why this happens 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.

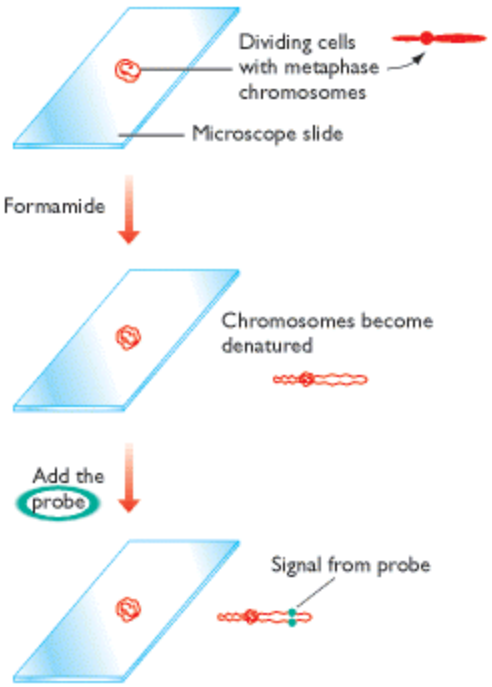


Figure 5.29. Fluorescent *in situ* hybridization. A sample of dividing cells is dried onto a microscope slide and treated with formamide so that the chromosomes become denatured but do not lose their characteristic metaphase morphologies (see Section 2.2.1). The position at which the probe hybridizes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

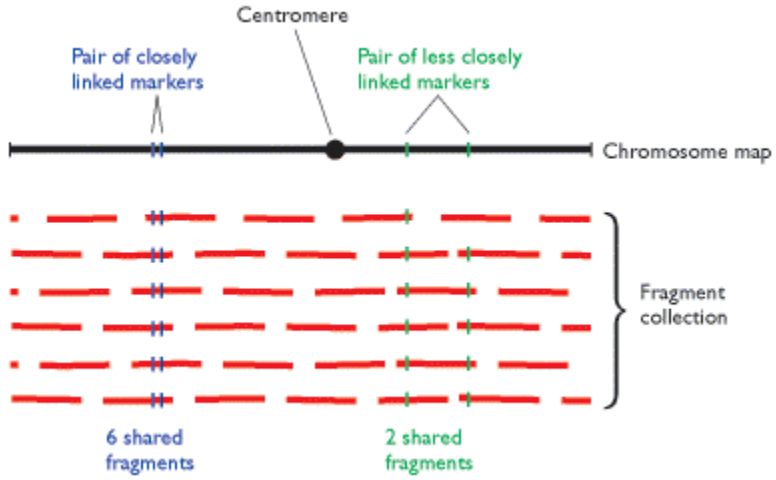


Figure 5.31. A fragment collection suitable for STS mapping. The fragments span the entire length of a chromosome, with each point on the chromosome present in an average of five fragments. The two blue markers are close together on the chromosome map and there is a high probability that they will be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

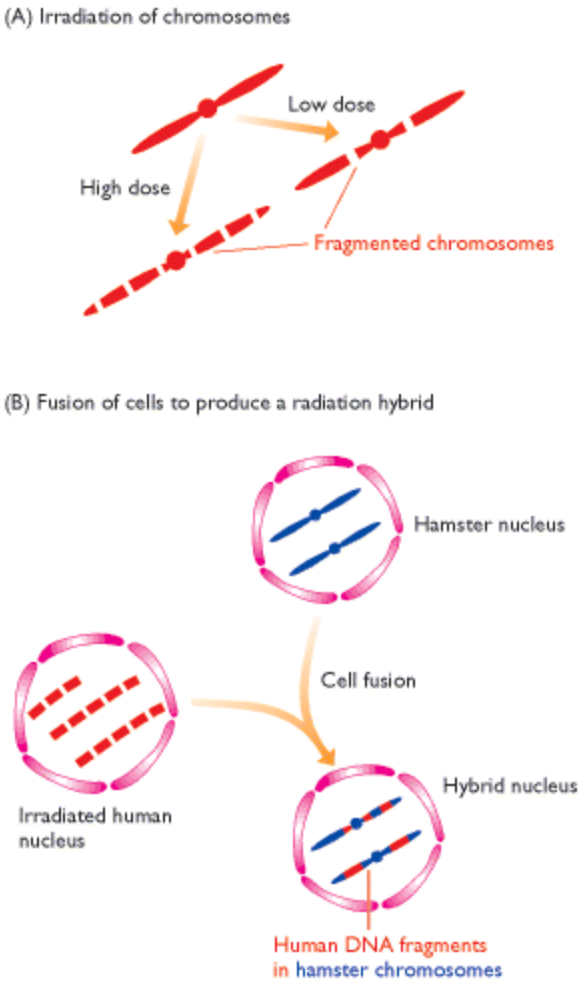


Figure 5.33. Radiation hybrids. (A) The result of irradiation of human cells: the chromosomes break into fragments, smaller fragments generated by higher X-ray doses. In (B), a radiation hybrid is produced by fusing an irradiated human cell with an untreated hamster cell. For clarity, only the nuclei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

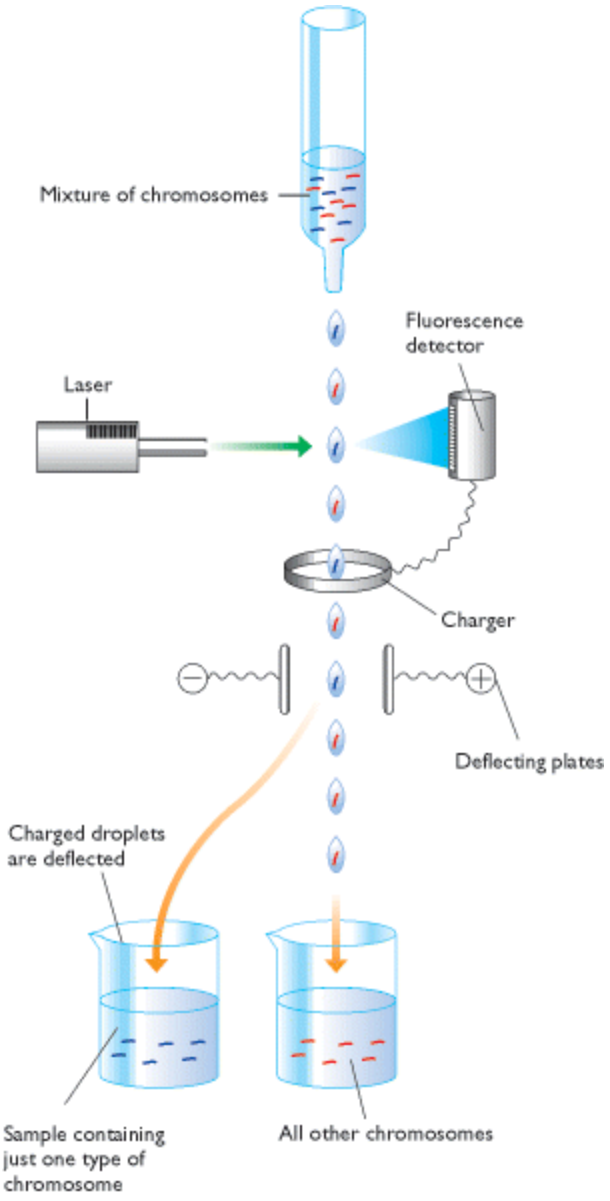


Figure 5.34. Separating chromosomes by flow cytometry. A mixture of fluorescently stained chromosomes is passed through a small aperture so that each drop that emerges contains just one chromosome. The fluorescence detector identifies the signal from drops containing the correct chromosome and applies 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.

cDNA libraries

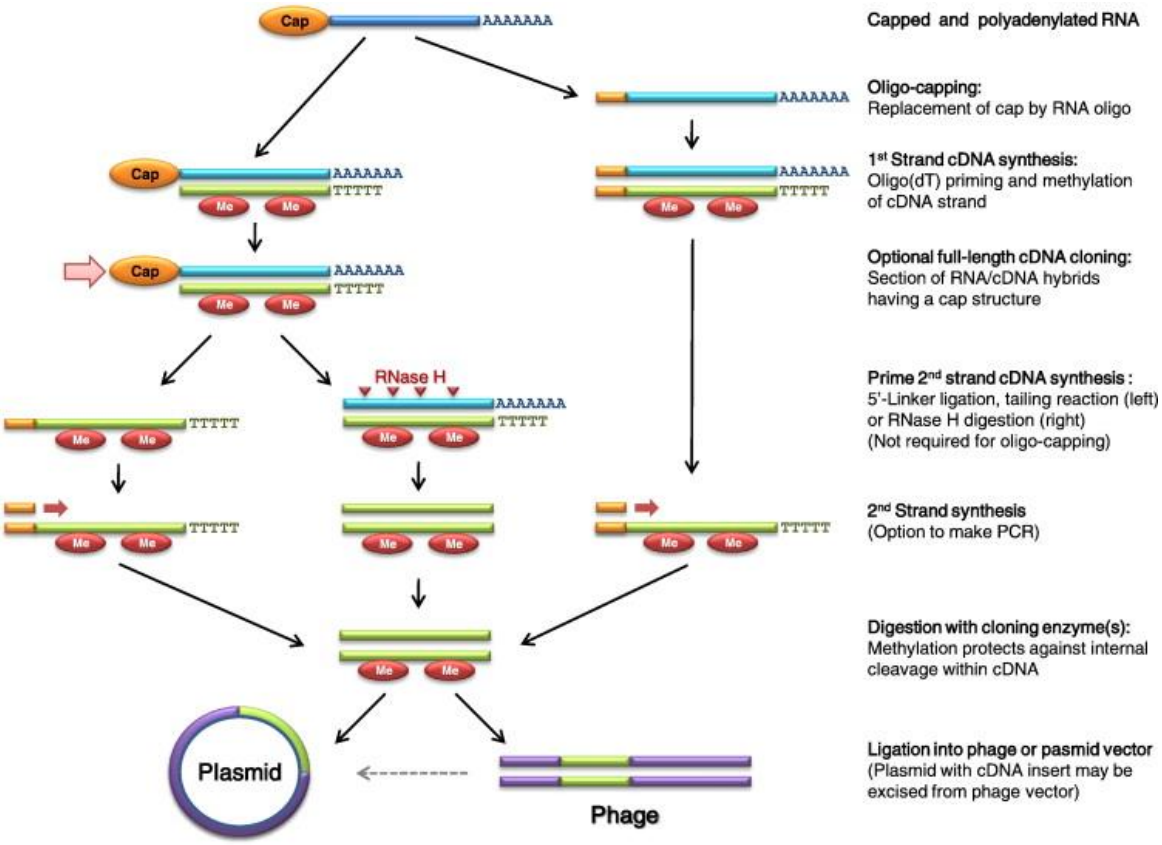
A **cDNA library** is a combination of cloned cDNA (complementary DNA) fragments inserted into a collection of host cells, which constitute some portion of the transcriptome of the organism and are stored as a "library". cDNA is produced from fully transcribed mRNA found in the nucleus and therefore contains only the expressed genes of an organism. Similarly, tissue-specific cDNA libraries can be produced. In eukaryotic cells the mature mRNA is already spliced, hence the cDNA produced lacks introns and can be readily expressed in a bacterial cell. While information in cDNA libraries is a powerful and useful tool since gene products are easily identified, the libraries lack information about enhancers, introns, and other regulatory elements found in a genomic DNA library.

Principle of cDNA cloning:

- **Complementary DNA (cDNA) cloning** is termed for the gene cloning (cloning of DNA fragments) obtained from cDNA.
- The principle of cDNA cloning is that it involves the copying of mRNA transcripts into DNA, which are then inserted into bacterial plasmids and then placed into bacteria by transformation.
- At this stage, it should be clear that mRNA used for cDNA preparation is a processed transcript and not the original one transcribed from DNA.
- In-order to clone a DNA sequence that codes for a required gene product, the gene should be removed from the organism and cloned it in the vector molecule.
- A gene library is a random collection of cloned fragments in an appropriate vector that particularly consists of all the genetic information about that species.
- There are two methods for the formation of gene libraries. They are:
    - **Complementary DNA (cDNA)**
    - **Genomic DNA libraries**

Steps of cDNA cloning:

1. **Isolation of mRNA**
2. **Synthesis of first strand of cDNA**
3. **Synthesis of second strand of cDNA**
4. **Cloning of cDNA**
5. **Introduction to host cells**
6. **Clone selection**



Capped and polyadenylated RNA

Oligo-capping:
Replacement of cap by RNA oligo

1st Strand cDNA synthesis:
Oligo(dT) priming and methylation of cDNA strand

Optional full-length cDNA cloning:
Section of RNA/cDNA hybrids having a cap structure

Prime 2nd strand cDNA synthesis :
5'-Linker ligation, tailing reaction (left) or RNase H digestion (right)
(Not required for oligo-capping)

2nd Strand synthesis
(Option to make PCR)

Digestion with cloning enzyme(s):
Methylation protects against internal cleavage within cDNA

Ligation into phage or pasmid vector
(Plasmid with cDNA insert may be excised from phage vector)

1. Isolation of mRNA:

- A crude extract of the tissue with the gene of interest is prepared.
- The extract must be free from proteins, polysaccharides and all other contaminants.
- The technique of oligo-deoxythymine (oligo-dT) cellulose chromatography is used for the further purification of many eukaryotic mRNAs from the total or polysomal fraction.
- mRNAs consist of poly A (adenosine residues) tail at their 3′ end.
- Under favourable conditions, this tail will bind to a string of thymidine residues immobilized on cellulose and then poly $(A)^+$ fraction can be eluted.
- Two or three passages of the poly $(A)^+$ fraction through such a column produces a fraction highly enriched for mRNA.
- This fraction includes different mRNA sequences, however certain techniques can be employed for extracting a particular mRNA species.
- After the preparation of the fraction, it is essential to confirm if the extracted mRNA consists of the sequence of interest.
- It is performed by translation of mRNA in vitro and identification of suitable polypeptides in the products obtained.

2. Synthesis of first strand of cDNA:

- Reverse transcriptase is a RNA dependent DNA polymerase which is used to copy the mRNA fraction into the first strand of DNA.
- This enzyme, like all other DNA polymerases, can only add residues at the 3′-OH group of an existing primer, which is base paired with the template.
- The most commonly used primer is oligo-dT for cloning of cDNAs.
- Oligo-dT primer is 12-18 nucleotides in length, that binds to the poly (A) tract at the 3′ end of mRNA molecules.
- The RNA strand of the resulting RNA-DNA hybrid is destroyed prior to second strand synthesis through alkaline hydrolysis.

3. Synthesis of second strand of cDNA:

- The second strand of cDNA can be synthesized by two techniques. They are:
- **i. Self-priming cDNA:**
    - In Self-priming, the mRNA hybrid obtained is denaturated for the synthesis of second strand on the single strand of cDNA by the klenow fragment of DNA polymerase I.
    - The transitory hairpin structure at the 3′ end of single-stranded DNA can be used to prime the synthesis of second strand of cDNA by the klenow fragment of *Escherichia coli* DNA polymerase I.
    - Single-strand specific S1 nuclease digests the hairpin loop and any single-stranded overhung at the other end.

- The ultimate product is a population of double-stranded, blunt-ended DNA molecules complementray to the original mRNA fraction.

- **ii. Replacement synthesis:**
  - In this method, the cDNA:mRNA hybrid works as a template for a nick translation reaction.
  - In the mRNA strand of the hybrid, RNase H produces nicks and gaps, creating a series of RNA primers.
  - These RNA primers are used by *E. coli* DNA polymerase I during the synthesis of second strand of cDNA.
  - The advantages of this technique are:
    - – very efficient
    - – can be performed directly using the products of the first strand reaction
    - – eliminates the need to use nuclease S1 to cleave the single-stranded hairpin loop in the double stranded cDNA.

4. Cloning of cDNA:

- The most frequently used technique for cloning cDNAs involves the addition of complementary homopolymeric tracts to double stranded cDNA and to the plasmid vector.
- To the cDNA, strings of cytosine residues are added using the enzyme terminal transferase to form oligo-dC tails on the 3′ ends.
- Likewise, a plasmid is cut open at a unique restriction endonuclease site and tailed with oligo-dG.
- Now, the vector and the double stranded cDNA are joined by hydrogen bonding between the complementary homopolymers.
- It results in the formation of open circular hybrid molecules capable of transforming *E. coli*.

5. Introduction to host cells:

- For the transforming of bacteria, the recombinant plasmids are used, usually the *E. coli* K-12 strain.
- The uptake of plasmid molecules from the surrounding medium is performed by E. coli cells treated with calcium chloride.
- Any gaps in the recombinant plasmid will be repaired by the host cells.
- The transformed bacteria can be isolated from non-transformed ones on the basis of antibiotic resistance.
- Majority of cloning plasmids contain two antibiotic resistance genes, one of which is destroyed during cloning.
- For instance, in the case of pBR322, cloning into unique PstI site destroys ampicillin resistance but leaves tetracycline resistance intact.

- Bacteria transformed with a recombinant plasmid will be sensitive to ampicillin but resistant to tetracycline.

6. Clone selection:

- The antibiotic resistance selection already performed has recognized which clones carry a recombinant plasmid, however there will be thousands of various inserts.
- The cloning process generally commences with a whole population of mRNA sequences.
- Selection of clones carrying the sequence of interest is the tough job.
- If the gene is expressed, then the simplest selection is to screen for the presence of the protein.
- It can be screened either by bacterial phenotype it produces or by the protein detection methods usually based on immunological or enzymological techniques.
- If the protein is not expressed, then other methods such as nucleic acid hybridization are used.
- Identification of the gene is discussed after the genomic DNA cloning.

Genomic DNA libraries

1.Genomic Libraries:

Genomic libraries are libraries of genomic DNA sequences. These can be produced using DNA from any organism.

2. Principle of Genomic Libraries:

A genomic library contains all the sequences present in the genome of an organism (apart from any sequences, such as telomeres that cannot be readily cloned). It is a collection of cloned, restriction-enzyme-digested DNA fragments containing at least one copy of every DNA sequence in a genome. The entire genome of an organism is represented as a set of DNA fragments inserted into a vector molecule.

3. Vectors used for the Construction of Genomic Library:

The choice of vectors for the construction of genomic library depends upon three parameters:

1. The size of the DNA insert that these vectors can accommodate.

- 2. The size of the library that is necessary to obtain a reasonably complete representation of the entire genome.
- 3. The total size of the genome of the target organism.
- In the case of organism with small genomic sizes, such as E. coli, a genomic library could be constructed by using a plasmid vector. In this case only 5000 clones (of average DNA insert size 5kb) would give a greater than 99% chance of cloning the entire genome (4.6 $x10^6$ bp).
- Most libraries from organisms with larger genomes are constructed using lambda phage, BAC or YAC vectors. These accept DNA inserts of approximately 23,45,350 and 1000kb

respectively. Due to this, fewer recombinants are needed for complete genome coverage in comparison to the use of plasmids.

4. Size of Genomic Library:

It is possible to calculate the number (N) of recombinants (plaques or colonies) that must be in a genomic library to give a particular probability of obtaining a given sequence.

The formula is:

N = In (1 – P)/ln (1 – f),

where 'P' is the desired probability and 'f is the fraction of the genome in one insert. For example, for a probability of 0.99 with insert sizes of 20kb this values for the E. coli (4.6 x 10^6 bp) and human (3 x 10^9 bp) genomes are:

$N_{g\ coli}$ = In (1 – 0.99) / In [1 – (2 x 10^4/4.6 x 10^6)] = 1.1 x 10^3

N_{human} = In (1 – 0.99)/ In [1 – (2 x 10^4/3 x 10^9)] = 6.9 x 10^5

These values explain why it is possible to make good genomic libraries from prokaryotes in plasmids where the insert size is 5-10 kb, as only a few thousand recombinants will be needed.

5. Types of Genomic Libraries:

Depending on the source of DNA used forced construction of genomic library it is of following two types:

(a) Nuclear Genomic Library:

This is genomic library which includes the total DNA content of the nucleus. While making such a library we specifically extract the nuclear DNA and use it for the making of the library.

(b) Organelle Genomic Library:

In this case we exclude the nuclear DNA and targets the total DNA of either mitochondria, chloroplast or both.

6. Procedure in the Construction of Genomic Library:

1. Preparing DNA:

The key to generating a high-quality library usually lies in the preparation of the insert DNA. The first step is the isolation of genomic DNA. The procedures vary widely according to the organism under study. Care should be taken to avoid physical damage to the DNA.

If the intention is to prepare a nuclear genomic library, then the DNA in the nucleus is isolated, ignoring whatever DNA is present in the mitochondria or chloroplasts. If the aim is to make an organelle genomic library, then it would be wise to purify the organelles away from the nuclei first and then prepare DNA from them.

2. Fragmentation of DNA:

The DNA is then fragmented to a suitable size for ligation into the vector. This could be done by complete digestion with a restriction endonuclease. But this has a demerit. Digestion by the use of restriction endonuclease produces DNA fragments which are not intact.

To solve this problem we use partial digestion with a frequently cutting enzyme (such as Sau3A, with a four-base-pair recognition site) to generate a random collection of fragments with a suitable size distribution.

Once prepared, the fragments that will form the inserts are often treated with phosphate, to remove terminal phosphate groups. This ensures that separate rate pieces of insert DNA cannot be ligated together before they are ligated into the vector. Ligation of separate fragments is undesirable, as it would generate clones containing non-contiguous DNA, and we would have no way of knowing where the joints lay

3. Vector Preparation:

This will depend on the kind of vector used. The vector needs to be digested with an enzyme appropriate to the insert material we are trying to clone.

4. Ligation and Introduction into the Host:

Vector and insert are mixed, ligated, packaged and introduced into the host by transformation, infection or' some other technique.

5. Amplification:

This is not always required. Libraries using phage cloning vectors are often kept as a stock of packaged phage. Samples of this can then be plated out on an appropriate host when needed. Libraries constructed in plasmid vectors are kept as collections of plasmid-containing cells, or as naked DNA that can be transformed into host cells when needed.

With storage, naked DNA may be degraded. Larger molecules are more likely to be degraded than smaller ones, so larger recombinants will be selectively lost, and the average insert size will fall.

7. Creation of a Genomic Library using the Phage-λ Vector EMBL3A:

High-molecular-weight genomic DNA is partially digested with Sau3Al. The fragments are treated with phosphatase to remove their 52 phosphate groups. The vector is digested with Bam/HI and EcoRI, which cut within the poly-linker sites.

The tiny BamHI/EcoRl poly-linker fragments are discarded in the iso-propanol precipitation, or alternatively the vector arms may be purified by preparative agarose gel electrophoresis. The vector arms are then ligated with the partially digested genomic DNA.

The phosphatase treatment prevents the genomic DNA fragments from ligating together. Non-recombinant vector cannot reform because the small poly-linker fragments have been discarded. The only package able molecules are recombinant phages. These are obtained as plaques on a P2 lysogen of sup+ E. coli. The Spi" selection ensures recovery of recombinant phage plaques.

8. Problems Associated with the Construction of Genomic Library:

In the making of a genomic library we digest the total genomic DNA with a restriction endonuclease, such as EcoRl, insert the fragments into a suitable phage X vector, and then attempt to isolate the desired clone. How many recombinants would we have to screen in order to isolate the right one?

Let us assume that EcoRI gives an average of about 4kb of DNA fragment, and given that the size of the human haploid genome is 2.8 x 106kb, it is clear that over 7 x 10^5 independent recombinants must be prepared and screened in order to obtain a desired sequence. In other words, we have to obtain a very large number of recombinants, which is a very labour intensive procedure.

There are three problems associated with the above approach:

1. The gene may be cut internally one or more times by Eco RI so that it is not obtained as a single fragment. This is likely if the gene is large.

2. Many times while making a library we want to obtain extensive regions flanking the gene or whole gene clusters. Fragments averaging about 4 kb are likely to be inconveniently short.

3. The obtained gene fragment may be larger than the size which the vector can accept. In this case the appropriate gene would not be cloned at all.

These problems can be overcome by cloning random DNA fragments of a large size. Since the DNA is randomly fragmented, there will be no exclusion of any DNA sequence. Also in this case the clones will overlap one another allowing the sequence of very large genes to be assembled. Because of the larger size of each cloned DNA fragment fewer clones are required for a complete or nearly complete library.

Now again we have a problem. How can appropriately sized random fragments be produced? Various methods are available of which random breakage by mechanical shearing is the most appropriate one. This is because the average fragment size can be controlled. Along with this the insertion of the resulting fragments into vectors requires additional modification steps.

A library representation of a eukaryotic organism would contain a very large number of clones, many of which would contain non-coding DNA such as repetitive DNA and regulatory regions. Also, eukaryotic genes often contain introns, which are un-translated regions interrupting the coding sequence.

These regions are normally copied into mRNA in the nucleus but spliced out before the mature mRNA is exported to the cytoplasm for translation into protein. Prokaryotic organisms are unable to do this processing so the mature mRNA cannot be made in E. coli and the protein will not be expressed.

If your screening method requires that the gene be expressed it will not work with a genomic library from a eukaryotic organism.

12. Applications of Genomic Library:

Genomic library has following applications:

1. It helps in the determination of the complete genome sequence of a given organism.

2. It serves as a source of genomic sequence for generation of transgenic animals through genetic engineering.

3. It helps in the study of the function of regulatory sequences in vitro.

4. It helps in the study of genetic mutations in cancer tissues.

5. Genomic library helps in identification of the novel pharmaceutical important genes.
6. It helps us in understanding the complexity of genomes.

UNIT –5- SBIA1301 – Molecular Biology and Genomics

Genome databases

Genomic Databases: contain data related to the genomic sequencing of different organisms, and gene annotations;

– Human Genome Databases: include information on the human gene sequencing.

– Model Organism Databases (MOD): store data coming from the sequencing projects of model organisms (such as, e.g., MATDB); they are also intended to support the Human Genome Project (HGP)

Other Organism Databases: store information derived from sequencing projects not related to HGP.

– Organelle Databases: store genomic data of cellular organelles, such as mitochondria, having their own genome, distinct from the nuclear genome.

– Virus Databases: store virus genomes.

Genomic databases contain a large set of data types. Some archives report only the sequence, the function and the organism corresponding to a given portion of genome, other ones contain also detailed information useful for biological or clinical analysis.

2.1 Recoverable Data

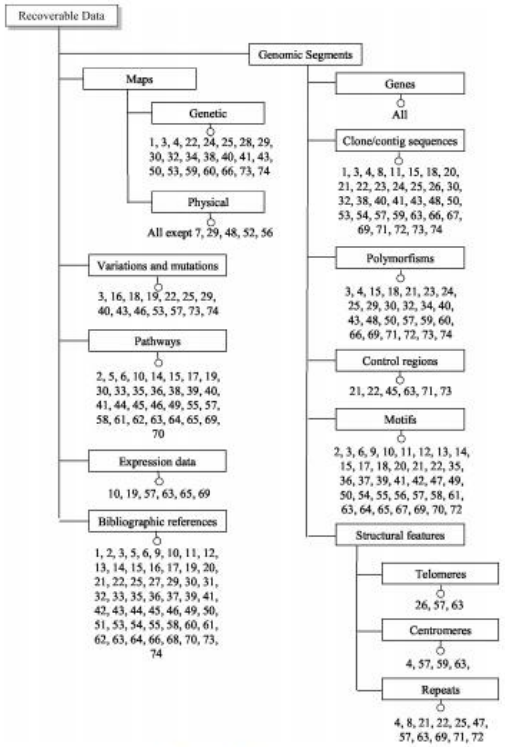


Fig. 2. Recoverable data typologies.

data that are recoverable from genomic databases can been distinguished in six classes: • Genomic segments: include all the nucleotide subsequences which are meaningful from a biological point of view, such as genes, clone/clontig sequences, polymorphisms, control regions, motifs and structural features of chromosomes. In detail:

- *Genes*: are DNA subsequences originating functional product such as proteins and regulatory elements. All genomic databases contain information about genes.
- *Clone/contig regions*: are sequences of in vitro copies of DNA regions, used to clone the DNA sequence of a given organism. Examples of databases storing information about clone/contig regions are ArkDB [1], SGD [32] and ZFIN [47].
- *Polymorphisms*: are frequent changes in nucleotide sequences, usually corresponding to a phenotypic change (e.g., the variety of eye colors in the population). As an example, BeetleBase [3] contains information about polymorphisms.
- *Control regions*: are DNA sequences regulating the gene expression. A few databases store information about control regions. Among them there are Ensembl [13], FlyBase [42], SGD [32], and WormBase [28].
- *Motifs (or patterns)*: are specific DNA segments having important functions and repeatedly occurring in the genome of a given organism. Examples of databases containing information about motifs are AureoList [2], Leproma [20] and VEGA [38].
- *Structural features*: can be further distinguished in *telomeres*, *centromeres* and *repeats*.
  - *Telomeres*: are the terminal regions of chromosomes. Génolevures [33], PlasmoDB [30] and SGD [32] store information about telomeres.
  - *Centromeres*: represent the conjunction between *short arms* and *long arms* of the chromosome. BGI-RISe [51], PlasmoDB [30], RAD [24] and SGD [32] are databases containing information about centromeres.
  - *Repeats*: are repetitions of short nucleotide segments, repeatedly occurring in some of the regions of the chromosome. Among the others, CADRE [7], PlasmoDB [30] and ToxoDB [37] store information about repeats.

- *Maps*: result from projects that produced the sequencing and mapping of the DNA of diverse organisms, such as the Human Genome Project [50]. In particular, *genetic maps* give information on the order in which genes occur in the genome, providing only an estimate of genes distances, whereas *physical maps* give more precise information on the physical distances among genes. GOBASE [45] and BeetleBase [3] are examples of databases storing maps.
- *Variations and mutations*: A nucleotide change is a mutation if it occurs with low frequency in the population (as opposed to polymorphisms). Mutations may cause alterations in the protein tertiary structures, inducing pathologic variations of the associated biological functions. These information are stored in databases such as, for example, FlyBase [42], and PlasmoDB [30].
- *Pathways*: describe interactions of sets of genes, or of proteins, or of metabolic reactions involved in the same biological function. Pathways are stored in Bovilist [5] and Leproma [20], for example.

• Expression data: are experimental data about the different levels of expression of genes. The levels of such an expression are related to the quantity of genic product of genes. Information about expression data are stored in, e. g., CandidaDB [6], PlasmoDB [30], SGD [32] and ToxoDB [37].

• Bibliographic references: are repositories of (sometimes, links to) relevant biological literature. Most genomic databases contain bibliographic references.

Genome annotation

DNA annotation or **genome annotation** is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.[1] Genes in eukaryotic genome can be annotated using FINDER.[2]

For DNA annotation, a previously unknown sequence representation of genetic material is enriched with information relating genomic position to intron-exon boundaries, regulatory sequences, repeats, gene names and protein products. This annotation is stored in genomic databases such as Mouse Genome Informatics, FlyBase, and WormBase. Educational materials on some aspects of biological annotation from the 2006 Gene Ontology annotation camp and similar events are available at the Gene Ontology website.[3]

The National Center for Biomedical Ontology (www.bioontology.org) develops tools for automated annotation[4] of database records based on the textual descriptions of those records.

As a general method, dcGO[5] has an automated procedure for statistically inferring associations between ontology terms and protein domains or combinations of domains from the existing gene/protein-level annotations.

Process

Genome annotation consists of three main steps:

1. identifying portions of the genome that do not code for proteins
2. identifying elements on the genome, a process called gene prediction
3. attaching biological information to these elements

Automatic annotation tools attempt to perform these steps via computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

A simple method of gene annotation relies on homology based search tools, like BLAST, to search for homologous genes in specific databases, the resulting information is then used to annotate genes and genomes.[7] However, as information is added to the annotation platform, manual annotators become capable of deconvoluting discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases (e.g. Ensembl) rely on curated data sources as well as a range of different software tools in their automated genome annotation pipeline.[8]

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- gene structure
- coding regions

- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and *in silico* analysis. Proteogenomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations.[9]

A variety of software tools have been developed to permit scientists to view and share genome annotations; for example, MAKER.

Genome annotation remains a major challenge for scientists investigating the human genome, now that the genome sequences of more than a thousand human individuals (The 100,000 Genomes Project, UK) and several model organisms are largely complete.[10][11] Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism.[7] Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together".[12]

Genome annotation is an active area of investigation and involves a number of different organizations in the life science community which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic means. Here is an alphabetical listing of on-going projects relevant to genome annotation:

- Encyclopedia of DNA elements (ENCODE)
- Entrez Gene
- Ensembl
- GENCODE
- Gene Ontology Consortium
- GeneRIF
- RefSeq
- Uniprot
- Vertebrate and Genome Annotation Project (Vega)

At Wikipedia, genome annotation has started to become automated under the auspices of the Gene Wiki portal which operates a bot that harvests gene data from research databases and creates gene stubs on that basis

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 universitiesand 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 byCharles 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 *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*	Finished genome sequences of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, plus whole-genome drafts of several others, including *C. briggsae*, *D. pseudoobscura*, 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 retroposons, 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.

Human genetic variation

Human genetic variation is the genetic differences in and among populations. There may be multiple variants of any given gene in the human population (alleles), a situation called polymorphism.

No two humans are genetically identical. Even monozygotic twins (who develop from one zygote) have infrequent genetic differences due to mutations occurring during development and gene copy-number variation.[1] Differences between individuals, even closely related individuals, are the key to techniques such as genetic fingerprinting. As of 2017, there are a total of 324 million known variants from sequenced human genomes.[2] As of 2015, the typical difference between an individual's genome and the reference genome was estimated at 20 million base pairs (or 0.6% of the total of 3.2 billion base pairs).[3]

Alleles occur at different frequencies in different human populations. Populations that are more geographically and ancestrally remote tend to differ more. The differences between populations represent a small proportion of overall human genetic variation. Populations also differ in the quantity of variation among their members. The greatest divergence between populations is found in sub-Saharan Africa, consistent with the recent African origin of non-African populations. Populations also vary in the proportion and locus of introgressed genes they received by archaic admixture both inside and outside of Africa.

The study of human genetic variation has evolutionary significance and medical applications. It can help scientists understand ancient human population migrations as well as how human groups are biologically related to one another. For medicine, study of human genetic variation may be important because some disease-causing alleles occur more often in people from specific geographic regions. New findings show that each human has on average 60 new mutations compared to their parents.

Causes of variation

Causes of differences between individuals include independent assortment, the exchange of genes (crossing over and recombination) during reproduction (through meiosis) and various mutational events.

There are at least three reasons why genetic variation exists between populations. Natural selection may confer an adaptive advantage to individuals in a specific environment if an allele provides a competitive advantage. Alleles under selection are likely to occur only in those geographic regions where they confer an advantage. A second important process is genetic drift, which is the effect of random changes in the gene pool, under conditions where most mutations are neutral (that is, they do not appear to have any positive or negative selective effect on the organism). Finally, small migrant populations have statistical differences - called the founder effect - from the overall populations where they originated; when these migrants settle new areas, their descendant population typically differs from their population of origin: different genes predominate and it is less genetically diverse.

In humans, the main cause is genetic drift. Serial founder effects and past small population size (increasing the likelihood of genetic drift) may have had an important influence in neutral differences between populations. The second main cause of genetic variation is due to the high degree of neutrality of most mutations. A small, but significant number of genes appear to have undergone recent natural selection, and these selective pressures are sometimes specific to one region

Measures of variation

Genetic variation among humans occurs on many scales, from gross alterations in the human karyotype to single nucleotide changes.[8] Chromosome abnormalities are detected in 1 of 160 live human births. Apart from sex chromosome disorders, most cases of aneuploidy result in death of the developing fetus (miscarriage); the most common extra autosomal chromosomes among live births are 21, 18 and 13.[9]

Nucleotide diversity is the average proportion of nucleotides that differ between two individuals. As of 2004, the human nucleotide diversity was estimated to be 0.1%[10] to 0.4% of base pairs.[11] In 2015, the 1000 Genomes Project, which sequenced one thousand individuals from 26 human populations, found that "a typical [individual] genome differs from the reference human genome at 4.1 million to 5.0 million sites … affecting 20 million bases of sequence"; the latter figure corresponds to 0.6% of total number of base pairs.[3] Nearly all (>99.9%) of these sites are small differences, either single nucleotide polymorphisms or brief insertions or deletions (indels) in the genetic sequence, but structural variations account for a greater number of base-pairs than the SNPs and indels.[3][12]

As of 2017, the Single Nucleotide Polymorphism Database (dbSNP), which lists SNP and other variants, listed 324 million variants found in sequenced human genomes

Single nucleotide polymorphisms

A single nucleotide polymorphism (SNP) is a difference in a single nucleotide between members of one species that occurs in at least 1% of the population. The 2,504 individuals characterized by the 1000 Genomes Project had 84.7 million SNPs among them.[3] SNPs are the most common type of sequence variation, estimated in 1998 to account for 90% of all sequence variants.[13] Other sequence variations are single base exchanges, deletions and insertions.[14] SNPs occur on average about every 100 to 300 bases[15] and so are the major source of heterogeneity.

A functional, or non-synonymous, SNP is one that affects some factor such as gene splicing or messenger RNA, and so causes a phenotypic difference between members of the species. About 3% to 5% of human SNPs are functional (see International HapMap Project). Neutral, or synonymous SNPs are still useful as genetic markers in genome-wide association studies, because of their sheer number and the stable inheritance over generations.[13]

A coding SNP is one that occurs inside a gene. There are 105 Human Reference SNPs that result in premature stop codons in 103 genes. This corresponds to 0.5% of coding SNPs. They occur

due to segmental duplication in the genome. These SNPs result in loss of protein, yet all these SNP alleles are common and are not purified in negative selection

Structural variations

Structural variation is the variation in structure of an organism's chromosome. Structural variations, such as copy-number variation and deletions, inversions, insertions and duplications, account for much more human genetic variation than single nucleotide diversity. This was concluded in 2007 from analysis of the diploid full sequences of the genomes of two humans: Craig Venter and James D. Watson. This added to the two haploid sequences which were amalgamations of sequences from many individuals, published by the Human Genome Project and Celera Genomics respectively.[17]

According to the 1000 Genomes Project, a typical human has 2,100 to 2,500 structural variations, which include approximately 1,000 large deletions, 160 copy-number variants, 915 Alu insertions, 128 L1 insertions, 51 SVA insertions, 4 NUMTs, and 10 inversions

Copy number variations

A copy-number variation (CNV) is a difference in the genome due to deleting or duplicating large regions of DNA on some chromosome. It is estimated that 0.4% of the genomes of unrelated humans differ with respect to copy number. When copy number variation is included, human-to-human genetic variation is estimated to be at least 0.5% (99.5% similarity). Copy number variations are inherited but can also arise during development.

A visual map with the regions with high genomic variation of the modern-human reference assembly relatively to a Neanderthal of 50k [26] has been built by Pratas et al.

Epigenetics

Epigenetic variation is variation in the chemical tags that attach to DNA and affect how genes get read. The tags, "called epigenetic markings, act as switches that control how genes can be read."[28] At some alleles, the epigenetic state of the DNA, and associated phenotype, can be inherited across generations of individuals.[29]

Genetic variability

Genetic variability is a measure of the tendency of individual genotypes in a population to vary (become different) from one another. Variability is different from genetic diversity, which is the amount of variation seen in a particular population. The variability of a trait is how much that trait tends to vary in response to environmental and genetic influences.

Clines

In biology, a cline is a continuum of species, populations, varieties, or forms of organisms that exhibit gradual phenotypic and/or genetic differences over a geographical area, typically as a result of environmental heterogeneity. In the scientific study of human genetic variation, a gene cline can be rigorously defined and subjected to quantitative metrics.

Haplogroups

In the study of molecular evolution, a haplogroup is a group of similar haplotypes that share a common ancestor with a single nucleotide polymorphism (SNP) mutation. The study of haplogroups provides information about ancestral origins dating back thousands of years.[33]

The most commonly studied human haplogroups are Y-chromosome (Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations. Y-DNA is passed solely along the patrilineal line, from father to son, while mtDNA is passed down the matrilineal line, from mother to both daughter or son. The Y-DNA and mtDNA may change by chance mutation at each generation.

Variable number tandem repeats

A variable number tandem repeat (VNTR) is the variation of length of a tandem repeat. A tandem repeat is the adjacent repetition of a short nucleotide sequence. Tandem repeats exist on many chromosomes, and their length varies between individuals. Each variant acts as an inherited allele, so they are used for personal or parental identification. Their analysis is useful in genetics and biology research, forensics, and DNA fingerprinting.

Short tandem repeats (about 5 base pairs) are called microsatellites, while longer ones are called minisatellites.

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 asantibody matrix) by Tse Wen Chang in 1983 in a scientific publication and a series of patents

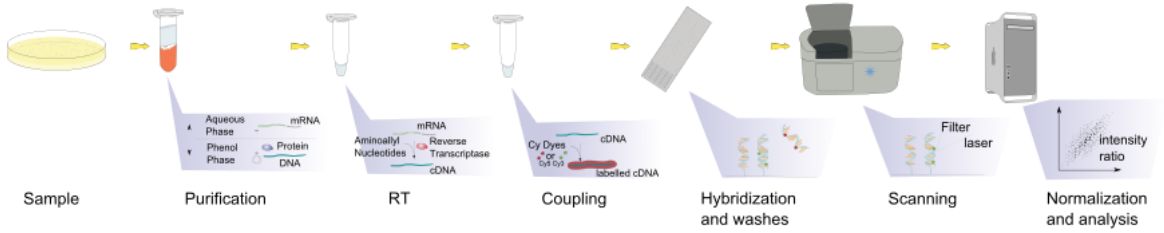
Types of microarrays include:

- 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 containspicomoles (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 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.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNAafter 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:

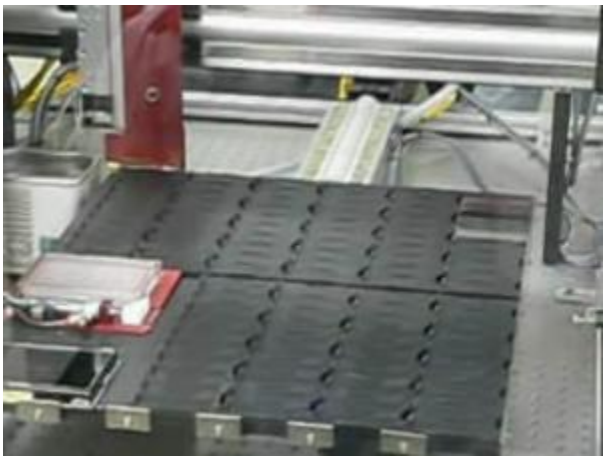
Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based 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]), mycoplasms 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 toimmunoprecipitate 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 splicingdetection	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 genesmicroarray	A Fusion gene microarray can detect fusion transcripts, e.g. from cancer specimens. The principle behind this is building on the alternative splicingmicroarrays. 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 theUniversity 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] orelectrochemistry on microelectrode arrays.

In spotted microarrays, the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond tomRNAs. 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]

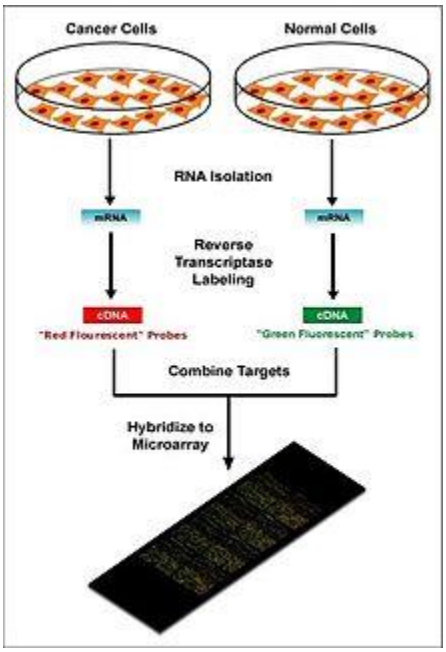


Diagram of typical dual-colourmicroarray experiment.

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.

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 Dual Chip platform for colorimetric Silver quant labeling, and Tele Chem 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.