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 – I -NEXT GENERATION SEQUENCING – SBI1606

High-throughput molecular analysis is a well-known technology that plays an important role in exploring biological questions in many species, especially in human genomic studies. Over the past 20 years, gene expression profiling, a revolutionary technique, has been widely used for genomic identification, genetic testing, drug discovery, and disease diagnosis, among other things (1). The field of genomics and proteomics research has undergone neoteric fluctuations as a result of next-generation sequencing (NGS), a paradigm-shifting technology that provides higher accuracy, larger throughput and more applications than the microarray platform (2-4). The use of massively parallel sequencing has increasingly been the object of study in recent years. The NGS technologies are implemented for several applications, including whole genome sequencing, de novo assembly sequencing, resequencing, and transcriptome sequencing at the DNA or RNA level.

Next-generation sequencing (NGS) refers to the deep, high-throughput, in-parallel DNA sequencing technologies developed a few decades after the Sanger DNA sequencing method first emerged in 1977 and then dominated for three decades . The NGS technologies are different from the Sanger method in that they provide massively parallel analysis, extremely high-throughput from multiple samples at much reduced cost. Millions to billions of DNA nucleotides can be sequenced in parallel, yielding substantially more throughput and mini- mizing the need for the fragment-cloning methods that were used with Sanger sequencing . The second-generation sequencing methods are characterized by the need to prepare amplified sequencing libraries before undertaking sequencing of the amplified DNA clones, whereas third-generation single molecular sequencing can be done without the need for creating the time-consuming and costly amplification libraries . The parallelization of a high number of sequencing reactions by NGS was achieved by the miniaturization of sequencing reactions and, in some cases, the development of microfluidics and improved detection systems. The time needed to generate the gigabase (Gb)-sized sequences by NGS was reduced from many years to only a few days or hours, with an accompanying massive price reduction.

For example, as part of the Human Genome Project, the J. C. Venter genome [7] took almost 15 years to sequence at a cost of more than 1 million dollars using the Sanger method, whereas the J. D. Watson (1962 Nobel Prize winner) genome was sequenced by NGS using the 454 Genome Sequencer FLX with about the same 7.5x coverage within 2 months and for approximately 100th of the price. The cost of sequencing the bacterial genome is now possible at about $1000 (https://www.nanoporetech.com), and the large-scale whole-genome sequencing (WGS) of 2,636 Icelanders has brought some of the aims of the 1000 Genomes Project to abrupt fruition Rapid progress in NGS technology and the simultaneous development of bioinformatics tools has allowed both small and large research groups to generate de novo draft genome sequences for any organism of interest. Apart from using NGS for WGS [11], these technologies can be used for whole transcriptome shotgun sequencing (WTSS) — also called RNA sequencing (RNA-seq) [12], whole-exome sequencing (WES) [13], targeted (TS) or candidate gene se- quencing (CGS) [14–16], and methylation sequencing (MeS) [17]. RNA-seq can be used to identify all transcriptional activities (coding and noncoding) or a select subset of targeted RNA transcripts within a

given sample, and it provides a more precise and sensitive measurement of gene expression levels than microarrays in the analysis of many samples. In contrast to WGS, WES provides coverage for more than 95% of human exons to investigate the protein-coding regions (CDS) of the genome and identify coding variants or SNPs when WGS and WTSS are not practical or necessary. Since the exome represents less than 2% of the human genome, it is the cost-effective alternative to WGS and RNA-seq in the study of human genetics and disease [13]. However, WGS may be preferred over WES because it provides more data with better uniformity of read coverage on disease-associated variants and reveals polymorphisms outside coding regions and genomic rearrangements. The analysis of the methylome by MeS complements WGS, WES, and CGS to determine the active methylation sites and the epigenetic markers that regulate gene expression, epistructural base variations, imprinting, development, differentiation, disease, and the epigenetic state. The impact of NGS technology is indeed egalitarian in that it allows both small and large research groups the possibility to provide answers and solutions to many different problems and questions in the fields of genetics and biology, including those in medicine, agriculture, forensic science, virology, microbiology, and marine and plant biology.

WHAT NGS DOES

• NGS provides a much cheaper and higherthroughput alternative to sequencing DNA than traditional Sanger sequencing. Whole small genomes can now be sequenced in a day.

• High-throughput sequencing of the human genome facilitates the discovery of genes and regulatory elements associated with disease.

• Targeted sequencing allows the identification of disease-causing mutations for diagnosis of pathological conditions.

• RNA-seq can provide information on the entire transcriptome of a sample in a single analysis without requiring previous knowledge of the genetic sequence of an organism. This technique offers a strong alternative to the use of microarrays in gene expression studies.

LIMITATIONS

• NGS, although much less costly in time and money in comparison to first-generation sequencing, is still too expensive for many labs. NGS platforms can cost more than $100,000 in start-up costs, and individual sequencing reactions can cost upward of $1,000 per genome.

• Inaccurate sequencing of homopolymer regions (spans of repeating nucleotides) on certain NGS platforms, including the Ion Torrent PGM, and short-sequencing read lengths (on average 200–500 nucleotides) can lead to sequence errors.

• Data analysis can be time-consuming and may require special knowledge of bioinformatics to garner accurate information from sequence data.

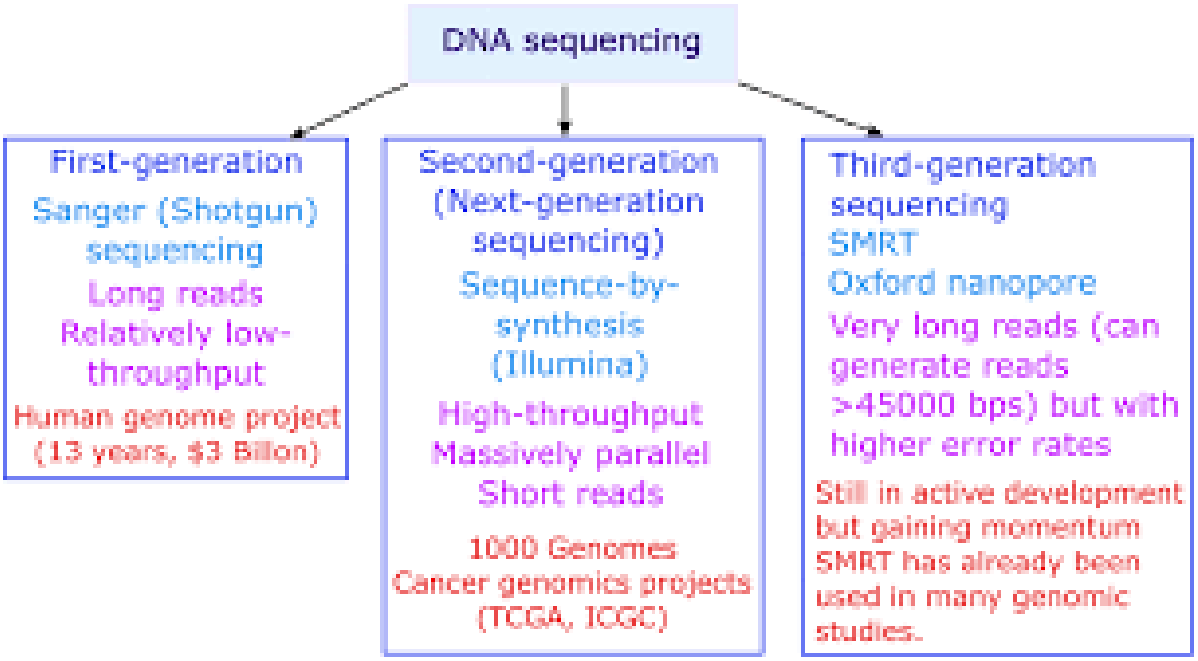
	Advantage	Disadvantage
Sanger	1. Long reading sequences Easy assembly in read outs (especially for GC rich highly repetitive DNA areas) 2. Smaller depth of sequencing required for good coverage 3. Easy to analyze 4. Relatively small data storage required	1. low Sensitivity (high allele frequency of cancer needed) 2. Scalable to few genes only 3. Unable to detect chromosomal aberrations 4. Insensitive to copy number alterations 5. High cost per base 6. Large amount of startup material required (1-3 ug) 7. Slower turnaround time
NGS	1. High sensitivity (tumor heterogeneity and stoma contamination will not be troubling) 2. High depth of sequencing is feasible 3. Scalable to entire genome 4. Detects chromosomal aberrations 5. Detects Copy number variations 6. Low cost per base 7. Small amount of startup material required (50 ng) 8. Quick turnaround time	1. Short reading sequences, challenges in assembly of the reads 2. Complicated data analysis 3. Large data storage required

Two generations of sequencing technology

Feature	First generation	Second generation
Isolate DNA fragments to sequence	Cloning in bacteria to generate many copies of the same DNA sequence, usually as a recombinant plasmid	Physical cloning to generate thousands of copies of a DNA molecule, separated on beads on or as positions on a flow cell
Purification of clones?	Prepare the plasmids from each bacterial clone	No need for plasmid preparation
DNA sequencing approach	Sequencing by synthesis or by base-specific degradation	Sequencing by synthesis, pyrosequencing, or ligation (SOLiD)
Method of detection	Electrophoresis to separate by size; fluorescent dyes	Light detection at each cycle of synthesis
Number of clones sequenced in parallel	scores to hundreds	hundreds of millions

1/20/14 3

Generations in sequencing



The First Generation of Sequencing

Sanger and Maxam-Gilbert sequencing technologies were classified as the First Generation Sequencing Technology [10,16] who initiated the field of DNA sequencing with their publication in 1977.

Sanger sequencing

Sanger Sequencing is known as the chain termination method or the dideoxynucleotide method or the sequencing by synthesis method. It consists in using one strand of the double stranded DNA as template to be sequenced. This sequencing is made using chemically modified nucleotides called dideoxy-nucleotides (dNTPs). These dNTPs are marked for each DNA bases by ddG, ddA, ddT, and ddC. The dideoxynucleotides are used dNTPs are used for elongation of nucleotide, once incorporated into the DNA strand they prevent the further elongation and the elongation is complete. Then, we obtain DNA fragments ended by a dNTP with different sizes. The fragments are separated according to their size using gel slab where the resultant bands corresponding to DNA fragments can be visualized by an imaging system (X-ray or UV light) [24,25]. Figure 1 details the Sanger sequencing technology. The first genomes sequenced by the Sanger sequencing are phiX174 genome with size of 5374 bp [26] and in 1980 the bacteriophage λ genome with length of 48501 bp [27]. After years of improvement, Applied Biosystems is the first company that has automated Sanger sequencing. Applied Biosystems has built in 1995 an automatic sequencing machine called ABI Prism 370 based on capillary electrophoresis allowing fast an accurate sequencing. The Sanger sequencing was used in several sequencing projects of different plant species such as Arabidopsis [28], rice [29] and soybean [30] and the most emblematic achievement of this sequencing technology is the decoding of the first human genome. The sanger sequencing was widely used for three decades and even today for single or low-throughput DNA

sequencing, however, it is difficult to further improve the speed of analysis that does not allow the sequencing of complex genomes such as the plant species genomes and the sequencing was still extremely expensive and time consuming. Maxam-Gilbert sequencing Maxam-Gilbert is another sequencing belonging to the first generation of sequencing known as the chemical degradation method. Relies on the cleaving of nucleotides by chemicals and is most effective with small nucleotides polymers. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of the four reactions (C, T+C, G, A+G). This reaction leads to a series of marked fragments that can be separated according to their size by electrophoresis. The sequencing here is performed without DNA cloning. However, the development and improvement of the Sanger sequencing method favored the latter to the Maxam-Gilbert sequencing method, and it is also considered dangerous because it uses toxic and radioactive chemicals.

The Second Generation of Sequencing

The first generation of sequencing was dominant for three decades especially Sanger sequencing, however, the cost and time was a major stumbling block. In 2005 and in subsequent years, have marked the emergence of a new generation of sequencers to break the limitations of the first generation. the basic characteristics of second generation sequencing technology are: (1) He generation of many millions of short reads in parallel, (2) He speed up of sequencing the process compared to the first generation, (3) He low cost of sequencing and (4) He sequencing output is directly detected without the need for electrophoresis. Short read sequencing approaches divided under two wide approaches: sequencing by ligation (SBL) and sequencing by synthesis (SBS), (more details for these sequencing categories are presented in [22,32]) and are mainly classified into three major sequencing platforms: Roche/454 launched in 2005, Illumina/Solexa in 2006 and in 2007 the ABI/SOLiD. We will briefly describe these commonly utilized sequencing platforms.

Roche/454 sequencing

Roche/454 sequencing appeared on the market in 2005, using pyrosequencing technique which is based on the detection of pyrophosphate released aier each nucleotide incorporation in the new synthetic DNA strand (http://www.454.com). He pyrosequencing technique is a sequencing-by-synthesis approach. DNA samples are randomly fragmented and each fragment is attached to a bead whose surface carries primers that have oligonucleotides complementary to the DNA fragments so each bead is associated with a single fragment (Figure 2A). Hen, each bead is isolated and amplified using PCR emulsion which produces about one million copies of each DNA fragment on the surface of the bead (Figure 2B). He beads are then transferred to a plate containing many wells called picotiter plate (PTP) and the pyrosequencing technique is applied which consists in activating of a series of downstream reactions producing light at each incorporation of nucleotide. By detecting the light emission after each incorporation of nucleotide, the sequence of the DNA fragment is

deduced (Figure 2C) [15]. tHe use of the picotiter plate allows hundreds of thousands of reactions occur in parallel, considerably increasing sequencing throughput [14]. tHe latest instrument launched by Roche/454 called GS FLX+ that generates reads with lengths of up to 1000 bp and can produce ~1Million reads per run (454.com GS FLX+Systems http//454.com/products/gs-flxsystem/index.asp). Other characteristics of Roche/454 instruments are listed in [16,25]. tHe Roche/454 is able to generate relatively long reads which are easier to map to a reference genome. He main errors detected of sequencing are insertions and deletions due to the presence of homopolymer regions [33,34]. Indeed, the identification of the size of homopolymers should be determined by the intensity of the light emitted by pyrosequencing. Signals with too high or too low intensity lead to under or overestimation of the number of nucleotides which causes errors of nucleotides identification.

Ion torrent sequencing

Life Technologies commercialized the Ion Torrent semiconductor sequencing technology in 2010 (https//www.thermofisher.com/us/en/ home/brands/ion-torrent.html). It is similar to 454 pyrosequencing technology but it does not use fluorescent labeled nucleotides like other second-generation technologies. It is based on the detection of the hydrogen ion released during the sequencing process [35]. 6pecifically, Ion Torrent uses a chip that contains a set of micro wells and each has a bead with several identical fragments. He incorporation of each nucleotide with a fragment in the pearl, a hydrogen ion is released which change the pH of the solution. His change is detected by a sensor attached to the bottom of the micro well and converted into a voltage signal which is proportional to the number of nucleotides incorporated (Figure 3). He Ion Torrent sequencers are capable of producing reads lengths of 200 bp, 400 bp and 600 bp with throughput that can reach 10 Gb for ion proton sequencer. He major advantages of this sequencing technology are focused on read lengths which are longer to other SGS sequencers and fast sequencing time between 2 and 8 hours. He major disadvantage is the diculty of interpreting the homopolymer sequences (more than 6 bp) [21,36] which causes insertion and deletion (indel) error with a rate about ~1%.

Illumina/Solexa sequencing

The Solexa company has developed a new method of sequencing. Illumina company (http://www.illumina.com) purchased Solexa that started to commercialize the sequencer Ilumina/Solexa Genome Analyzer (GA) [3,37]. Illumina technology is sequencing by synthesis approach and is currently the most used technology in the NGS market. THe sequencing process is shown in Figure 4. During the first step, the DNA samples are randomly fragmented into sequences and adapters are ligated to both ends of each sequence. THen, these adapters are fixed themselves to the respective complementary adapters, the latter are hooked on a slide with many variants of adapters (complementary) placed on a solid plate (Figure 4A). During the second step, each attached sequence to the solid plate is amplified by "PCR bridge amplification" that creates several identical copies of each sequence; a set of sequences made from the same original sequence is called a cluster. Each

cluster contains approximately one million copies of the same original sequence (Figure 4B). He last step is to determine each nucleotide in the sequences, Illumina uses the sequencing by synthesis approach that employs reversible terminators [38] in which the four modified nucleotides, sequencing primers and DNA polymerases are added as a mix, and the primers are hybridized to the sequences. THen, polymerases are used to extend the primers using the modified nucleotides. Each type of nucleotide is labeled with a fluorescent specific in order for each type to be unique. THe nucleotides have an inactive 3'-hydroxyl group which ensures that only one nucleotide is incorporated. Clusters are excited by laser for emitting a light signal specific to each nucleotide, which will be detected by a coupled-charge device (CCD) camera and Computer programs will translate these signals into a nucleotide sequence (Figure 4C). He process continues with the elimination of the terminator with the fluorescent label and the starting of a new cycle with a new incorporation [21,39]. He first sequencers Illumina/Solexa GA has been able to produce very short reads ~35 bp and they had an advantage in that they could produce paired-end (PE) short reads, in which the sequence at both ends of each DNA cluster is recorded. He output data of the last Illumina sequencers is currently higher than 600 Gpb and lengths of short reads are about 125 bp. Details on Illumina sequencers [13]. One of the main drawbacks of the Illumina/Solexa platform is the high requirement for sample loading control because overloading can result in overlapping clusters and poor sequencing quality. THe overall error rate of this sequencing technology is about 1%. Substitutions of nucleotides are the most common type of errors in this technology [40], the main source of error is due to the bad identification of the incorporated nucleotide.

ABI/SOLiD sequencing

Supported Oligonucleotide Ligation and Detection (SOLiD) is a NGS sequencer Marketed by Life Technologies (http:// www.lifetechnologies.com). In 2007, Applied Biosystems (ABI) has acquired SOLiD and developed ABI/SOLID sequencing technology that adopts by ligation (SBL) approach [3]. THe ABI/SOLiD process consists of multiple sequencing rounds. It starts by attaching adapters to the DNA fragments, fixed on beads and cloned by PCR emulsion. THese beads are then placed on a glass slide and the 8-mer with a fluorescent label at the end are sequentially ligated to DNA fragments, and the color emitted by the label is recorded (Figure 5A). THen, the output format is color space which is the encoded form of the nucleotide where four fluorescent colors are used to represent 16 possible combinations of two bases. He sequencer repeats this ligation cycle and each cycle the complementary strand is removed and a new sequencing cycle starts at the position n-1 of the template. THe cycle is repeated until each base is sequenced twice (Figure 5B). He recovered data from the color space can be translated to letters of DNA bases and the sequence of the DNA fragment can be deduced [15]. ABI/SOLiD launched the first sequencer that produce short reads with length 35 bp and output of 3 Gb/run and continued to improve their sequencing which increased the length of reads to 75 bp with an output up to 30 Gb/run [22,23]. THe strength of ABI/SOLiD platform is high accuracy because each base is read twice while the drawback is the relatively short reads and long run times. THe errors of sequencing in this technology is

due to noise during the ligation cycle which causes error identification of bases. THe main type of error is substitution.

The third Generation of Sequencing

The second-generation of sequencing technologies previously discussed have revolutionized the analysis of DNA and have been the most widely used compared to the first generation of sequencing technologies. However, the SGS technologies generally require PCR amplification step which is a long procedure in execution time and expansive in sequencing price. Also, it became clear that the genomes are very complex with many repetitive areas that SGS technologies are incapable to solve them and the relatively short reads made genome assembly more diffcult. To remedy the problems caused by SGS technologies, scientists have developed a new generation of sequencing called "third generation sequencing". Hese third generations of sequencing have the ability to offer a low sequencing cost and easy sample preparation without the need PCR amplification in an execution time significantly faster than SGS technologies. In addition, TGS are able to produce long reads exceeding several kilobases for the resolution of the assembly problem and repetitive regions of complex genomes. THere are two main approaches that characterize TGS [22]: He single molecule real time sequencing approach (SMRT) [38] that was developed by Quake laboratory [41-43] and the synthetic approach that rely on existing short reads technologies used by Illumina (Moleculo) [43] and 10xGenomics (https://www.10xgenomics.com) to construct long reads. THe most widely used TGS technology approach is SMRT and the sequencers that have used this approach are Pacific Biosciences and Oxford Nanopore sequencing (specifically the MinION sequencer). In the following, we present the two most widely used sequencing platforms in TGS to know Pacific Biosciences and the MinION sequencing from Oxford Nanopore technology.

Pacific biosciences SMRT sequencing

Pacific Biosciences (http//www.pacificbiosciences.com/) developed the first genomic sequencer using SMRT approach and it's the most widely used third-generation sequencing technology. Pacific Biosciences uses the same fluorescent labelling as the other technologies, but instead of executing cycles of amplification nucleotide, it detects the signals in real time, as they are emitted when the incorporations occur. It uses a structure composed of many SMRT cells, each cell contains microfabricated nanostructures called zeromode waveguides (ZMWs) which are wells of tens of nanometers in diameter microfabricated in a metal film which is in turn deposited onto a glass substrate [44,45]. These ZMWs exploit the properties of light passing through openings with a diameter less than its wavelength, so light cannot be propagated. Due to their small diameter, the light intensity decreases along the wells and the bottom of the wells illuminated (Figure 6A). Each ZMW contains a DNA polymerase attached to their bottom and the target DNA fragment for sequencing. During the sequencing reaction, the DNA fragment is incorporated by the DNA polymerase with fluorescent labelled

nucleotides (with different colors). Whenever a nucleotide is incorporated, it releases a luminous signal that is recorded by sensors (Figure 6B). The detection of the labelled nucleotides makes it possible to determine the DNA sequence. Compared to SGS, Pacific Bioscience technology has several advantages. THe preparation of the sample is very fast, it takes 4 to 6 hours instead of days [16]. In addition, the long-read lengths, currently averaging ~10 kbp [46] but individual very long reads can be as long as 60 kbp, which is longer than that of any SGS technology. Pacific Biosciences sequencing platforms have a high error rate of about 13% [13] dominated by insertions and deletions errors. These errors are randomly distributed along the long read.

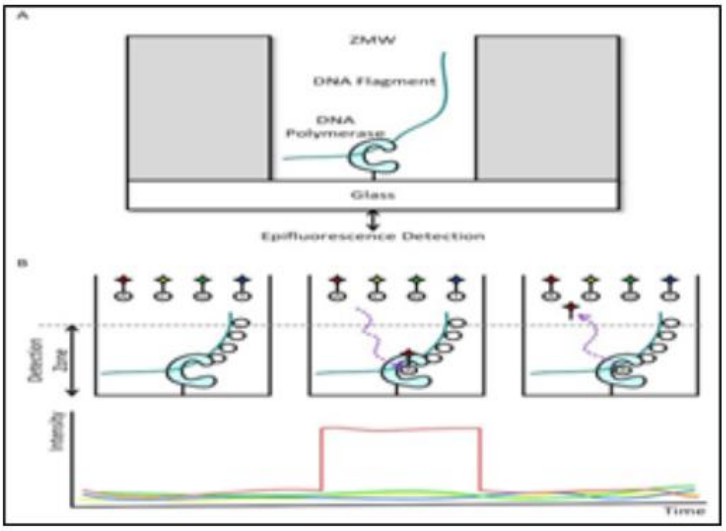


Figure 6: Pacific biosciences sequencing technology [39].

Oxford nanopore sequencing

THe Oxford Nanopore sequencing (ONT) was developed as a technique to determine the order of nucleotides in a DNA sequence. In 2014, Oxford Nanopore Technologies released the MinION [48] device that promises to generate longer reads that will ensure a better resolution structural genomic variants and repeat content [49]. It's a mobile single-molecule Nanopore sequencing measures four inches in length and is connected by a USB 3.0 port of a laptop computer. THis device has been released for testing by a community of users as part of the MinION Access Program (MAP) to examine the performance of the MinION sequencer [50]. In this sequencing technology, the first strand of a DNA molecule is linked by a hairpin to its complementary strand. THe DNA fragment is passed through a protein nanopore (a nanopore is a nanoscale hole made of proteins or synthetic materials [39]). When the DNA fragment is translated through the pore by the action of a motor protein attached to the pore, it generates a variation of an ionic current caused by differences in the moving nucleotides occupying the pore (Figure 7A). THis variation of ionic current is recorded progressively on a graphic model and then interpreted to identify the sequence (Figure 7B). THe sequencing is made on the direct strand generating the "template read" and then the hairpin structure is read followed by the inverse strand generating the "complement read", these reads is called "1D". If the "temple" and "complement" reads are combined, then we have a resulting consensus sequence called "two direction read" or "2D" [51,52]. Among the advantages offered by this sequencer: first, it's low cost and small size. Hen, the sample is loaded into a port on the device and data is displayed on the screen and generated without having to wait till the run is complete. And, MinION can provide very long reads exceeding 150 kbp which can improve the contiguity of the denovo assembly. However, MinION produces a high error rate of ~12% distributed about ~3% mismatchs, ~4% insertions and ~5% deletions [53]. THe ONT technology has continued to evolve. Recently, a new instrument has emerged called "PromethION"[54]; it is the bigger brother of the MinION [55]. It is an autonomous

worktable sequencer with 48 individual flow cells each with 3000 pores (equivalent to 48 MinIONs) operating at 500 bp [51] per second which is sufficiently powerful to achieve an ultra-high throughput needed for sequencing large genomes such as the human genome. Although the PromethION is not commercially available, the ONT announces that it is capable of producing ~2 to 4 Tb for a duration of 2 days and a length of reads [22] which can attain 200 Kpb which puts this sequencer in competition with the PacBioRSII sequencer from pacific biosciences in terms of read length and HiSeq sequencer from Illumina in cost.

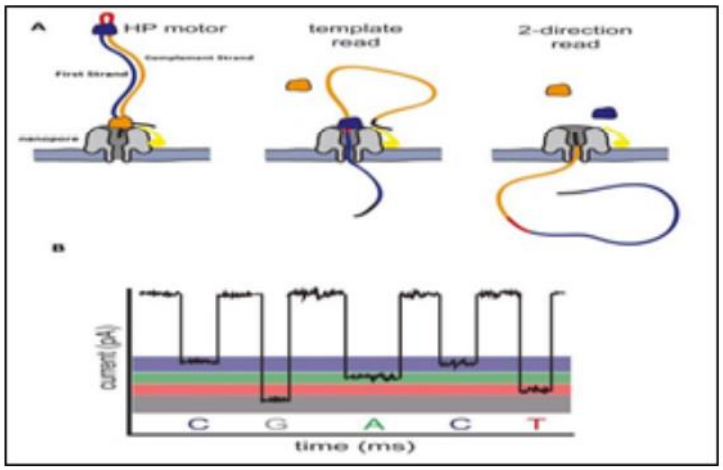
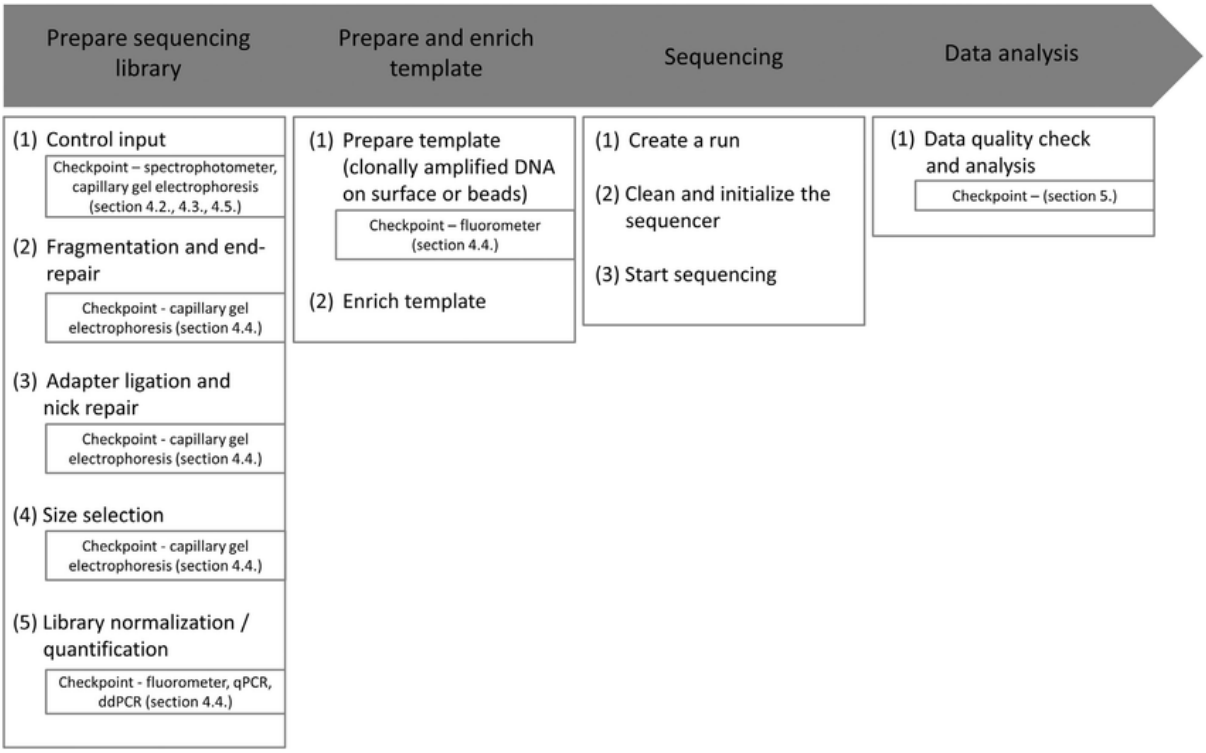


Figure 7: Oxford nanopore MinION sequencing [21].

NGS workflow



NGS Library

In NGS, a library is defined as a collection of DNA/RNA fragments that represents either the entire genome/transcriptome or a target region. Each NGS platform has its specificities, but, in simple terms, the preparation of an NGS library starts with the fragmentation of the starting material, then sequence adaptors are connected to fragments to allow the enrichment of those fragments. A good library should have great sensitivity and specificity. This means that all fragments of interest should be equally represented in the library and should not contain random errors (non-specific products). However, it is easier said than done, as genomic regions are not equally prone to be sequenced, making the construction of a sensitive and specific library challenging [10].

The first step to prepare libraries in most NGS workflows is the fragmentation of nucleic acid. Fragmentation can be done either by physical or enzymatic methods [11,12]. Physical methods include acoustic shearing, sonication and hydrodynamic shear. The enzymatic methods include digestion by DNase I or Fragmentase. Knierim and co-works, compared both enzymatic and physical fragmentation methods and found similar yields, showing that the choice between physical or enzymatic method only relies on experimental design or external factors, such as lab facilities [13].

Once the starting DNA has been fragmented, adaptors are connected to those fragments. The adaptors are introduced to create known begins and ends to random sequences allowing the sequencing process. An alternative strategy was developed that combines fragmentation and adaptor ligation in a single step, thus making the process simpler, faster and requiring a reduce sample input. The process is known as tagmentation and is based on transposon-based technology [14]. Upon nucleic acid fragmentation, the fragments are select according to the desired library size. This is limited either by the type of NGS instrument and by the specific sequencing application.

Short-read sequencers, such as Illumina and Ion Torrent, present best results when DNA libraries contain shorter fragments of similar sizes. Illumina fragments are longer than in Ion Torrent and can go up to 1500 bases in length [11] while in Ion Torrent the fragments can go up to 400 bases in length [15]. In contrast, long-read sequencers, like PacBio RS II [16] tend to produce ultra-long reads by fully sequencing a DNA fragment. The optimal library size is also limited by the sequencing application. For whole-genome sequencing, the longer fragments are preferable, while for RNA-seq and exome sequencing smaller fragments are feasible since most of the human exons are under 200 base pairs in length [17].

Next, an enrichment step is required, where the amount of target material is increased in a library to be sequenced. When just a part of the genome needs to be investigated both for research or clinical applications, it is known as target libraries. Basically, two methods are commonly used for such targeted approaches: capture hybridization-based sequencing and amplicon-based sequencing [18,19]. In the hybrid capture method, upon the fragmentation step, the fragmented molecules are hybridized specifically to DNA fragments complementary to the targeted regions of interest. This could be done by different methods such as

microarray technology or using biotinylated oligonucleotide probes [20], which aims to physically capture and isolate the sequences of interest. Two well-known examples of commercial library prep solutions based on hybrid capture methods are the SureSelect (Agilent Technologies) and SeqCap (Roche). Concerning the amplicon-based methods, those are based on the design of synthetic oligonucleotides (or probes), with a complementary sequence to the flanking regions of the target DNA to be sequenced. HaloPlex (Agilent Technologies) and AmpliSeq (Ion Torrent) are two examples of commercial library prep solutions based on amplicon-based strategies.

The amplicon-based methods have the limitations intrinsic to PCR amplifications, such as bias, PCR duplicates, primer competition and non-uniform amplification of target regions (due to variation in GC content) [21]. Hybrid capture methods were shown to be superior to amplicon-based methods, providing much more uniform coverage and depth than amplicon assays [19]. However, hybridization methods have the drawback of higher costs due to the specificity of the method (cost of the probes, experimental design, software, etc.) and are more time consuming than amplicon approaches. Hence, several attempts have been performed to overcome PCR limitations. One promising strategy is the Unique molecular identifiers (UMIs) that are short DNA molecules, which are ligated to library fragments [22]. Those UMI have a random sequence composition that assures that every fragment with a UMI is unique in your library. This allows that after PCR enrichment, PCR duplicates can be found by searching for non-unique fragment-UMI combinations, while the real biological duplicated will contain those UMI sequences [23,24].

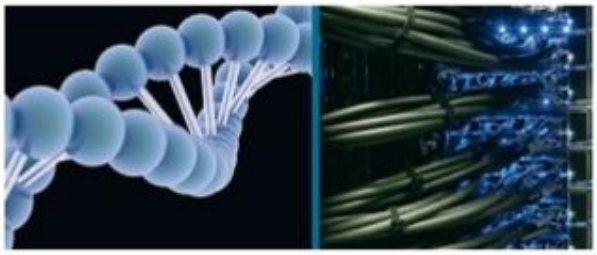
Applications of NGS

NGS technologies have many applications such as DNA-sequencing and assembly to determine an unknown genome without any preparation or search for variations among genome samples, RNA-sequencing [26,27], to analyze gene expression [28] and to predominantly identify DNA regions of DNA binding proteins, for example, transcription factors etc. The most important application of NGS is in identifying mutations. Commonly, short i.e. 50–250 bp NGS reads are initially mapped to a reference genome and after that from the mapped data, variations are detected. While most of the NGS applications concentrate on identification of single nucleotide variations (SNVs) or small insertions/ deletions (indels), structural variation including translocations, bigger indels, and copy number variation (CNV) can also be recognized from similar data. Structural variation discovery can be performed from whole genome NGS data or “targeted” data including exomes or gene panels. While targeted sequencing incredibly increments sequencing coverage or depth of specific genes, it might present predispositions in the data that require particular computational analysis. Since the past few years, there have been extensive advances in methods used to identify structural variations and a full coverage of variations from SNVs; balanced translocations to CNV can now be identified with reasonable sensitivity from either whole genome or targeted NGS data. Such methods are connected to clinical testing where they can supplement fluorescence in situ hybridization or array-based

testing. The identification of structural DNA variation has since quite a while ago assumed a part in the diagnosis of cancer and Mendelian disorders, originating before the approach of current DNA sequencing [29,30]. Structural DNA variation is found in a DNA region larger than 1 kb and incorporates a few classes, for example, translocations, inversions, insertions/deletions (indels) and copy number variations (CNVs) [31]. NGS-based diagnostics implement some portion of the clinical genomic testing in which a limited set of genes are targeted and not the entire genome and exome. Such diagnostics are routinely offered by more than 250 commercial and academic laboratories. One of the key elements of NGS-based diagnostics is its capacity to identify a full coverage of hereditary variation, offering the possibility to significantly streamline testing by utilizing a single analysis platform. For instance, prognostic assessment of acute myeloid leukemia for the most part requires the utilization of various advances including PCR and fragment sizing to detect FLT3 internal tandem duplications and NPM1 insertions, Sanger sequencing to identify CEBPA, IDH1/2, and DNMT3A mutations, and FISH to identify MLL, RARA, CBFB, and RUNX1 rearrangements. Such complicated assessments require very well trained staff with prohibitive cost. Thus, NGS-based testing can identify SNVs, insertions and translocations in a single test, considerably bringing down cost as compared with that of a conventional workup [32,33]. Single-cell sequencing is used for characterization of cancer heterogeneity. Cancer heterogeneity is caused due to different factors such as tissue hierarchies, clonal evolution, rare cells and dynamic cell states. With single-cell sequencing, it can be characterized in a large population of cells and molecular properties influencing clinical outcomes like prognosis and treatment, and can be determined in contrast to bulk sequencing, in which significant information is lost as the molecular profile represents an average phenotype over a large number of cells [34].

APPLICATIONS OF NGS

- Mutation discovery
- Transcriptome Analysis – RNA-Seq
- Sequencing clinical isolates in strain-to-reference mechanisms.
- Enabling Metagenomics
- Defining DNA-Protein interactions – ChIP-Seq
- Discovering non-coding RNAs
- Molecular diagnostics for Oncology & Inherited Disease study.
- Gene Regulation Analysis
- Whole Genome Sequencing
- Exploring Chromatin Packaging



NGS WORK FLOW

Next generation methods of DNA sequencing have three general steps:

- Library preparation: libraries are created using random fragmentation of DNA, followed by ligation with custom linkers
- Amplification: the library is amplified using clonal amplification methods and PCR
- Sequencing: DNA is sequenced using one of several different approaches

Step 1 in NGS Workflow: Library Prep

Library preparation is crucial to the success of your NGS workflow. This step prepares DNA or RNA samples to be compatible with a sequencer. Sequencing libraries are typically created by fragmenting DNA and adding specialized adapters to both ends. In the Illumina sequencing workflow, these adapters contain complementary sequences that allow the DNA fragments to bind to the flow cell. Fragments can then be amplified and purified.

To save resources, multiple libraries can be pooled together and sequenced in the same run—a process known as multiplexing. During adapter ligation, unique index sequences, or "barcodes," are added to each library. These barcodes are used to distinguish between the libraries during data analysis.

Step 2 in NGS Workflow: Sequencing

During the sequencing step of the NGS workflow, libraries are loaded onto a flow cell and placed on the sequencer. The clusters of DNA fragments are amplified in a process called cluster generation, resulting in millions of copies of single-stranded DNA. On most Illumina sequencing instruments, clustering occurs automatically.

In a process called sequencing by synthesis (SBS), chemically modified nucleotides bind to the DNA template strand through natural complementarity. Each nucleotide contains a fluorescent tag and a reversible terminator that blocks incorporation of the next base. The fluorescent signal indicates which nucleotide has been added, and the terminator is cleaved so the next base can bind.

After reading the forward DNA strand, the reads are washed away, and the process repeats for the reverse strand. This method is called paired-end sequencing.

Step 3 in NGS Workflow: Data Analysis

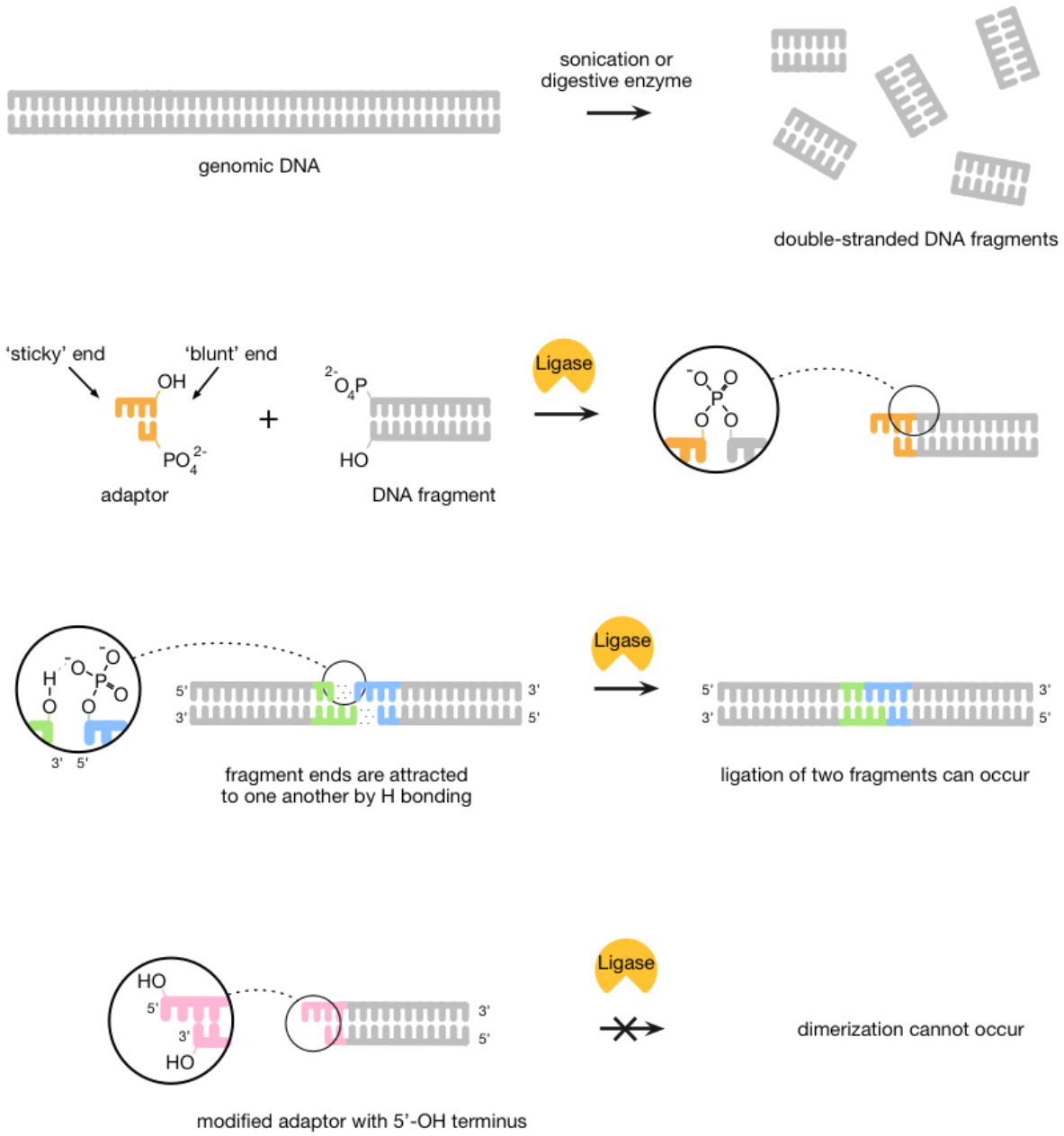
After sequencing, the instrument software identifies nucleotides (a process called base calling) and the predicted accuracy of those base calls. During data analysis, you can import your sequencing data into a standard analysis tool or set up your own pipeline.

Today, you can use intuitive data analysis apps to analyze NGS data without bioinformatics training or additional lab staff. These tools provide sequence alignment, variant calling, data visualization, or interpretation.

LIBRARY PREPARATION

Firstly, DNA is fragmented either enzymatically or by sonication (excitation using ultrasound) to create smaller strands. Adaptors (short, double-stranded pieces of synthetic DNA) are then ligated to these fragments with the help of DNA ligase, an enzyme that joins DNA strands. The adaptors enable the sequence to become bound to a complementary counterpart.

Adaptors are synthesised so that one end is 'sticky' whilst the other is 'blunt' (non-cohesive) with the view to joining the blunt end to the blunt ended DNA. This could lead to the potential problem of base pairing between molecules and therefore dimer formation. To prevent this, the chemical structure of DNA is utilised, since ligation takes place between the 3′-OH and 5′-P ends. By removing the phosphate from the sticky end of the adaptor and therefore creating a 5′-OH end instead, the DNA ligase is unable to form a bridge between the two termini (Figure 1).



In order for sequencing to be successful, the library fragments need to be spatially clustered in PCR colonies or 'polonies' as they are conventionally known, which consist of many copies of a particular library fragment. Since these polonies are attached in a planar fashion, the features of the array can be manipulated enzymatically in parallel. This method of library construction is much faster than the previous labour intensive procedure of colony picking and E. coli cloning used to isolate and amplify DNA for Sanger sequencing, however, this is at the expense of read length of the fragments.

AMPLIFICATION

Library amplification is required so that the received signal from the sequencer is strong enough to be detected accurately. With enzymatic amplification, phenomena such as 'biasing' and 'duplication' can occur leading to preferential amplification of certain library fragments.

Instead, there are several types of amplification process which use PCR to create large numbers of DNA clusters.

Emulsion PCR

Emulsion oil, beads, PCR mix and the library DNA are mixed to form an emulsion which leads to the formation of micro wells (Figure 2).

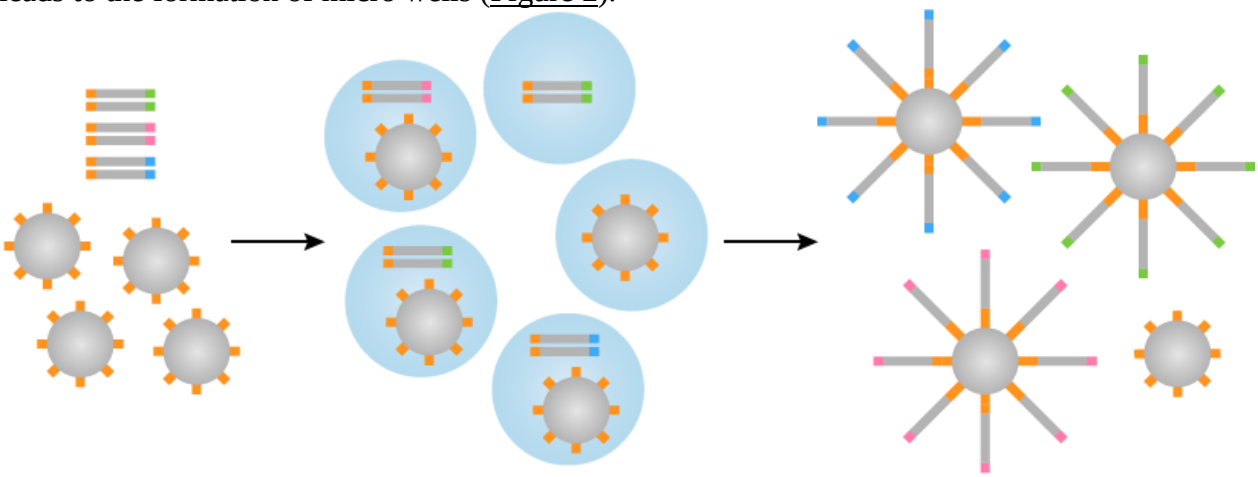


Figure 2 | Emulsion PCR

In order for the sequencing process to be successful, each micro well should contain one bead with one strand of DNA (approximately 15% of micro wells are of this composition). The PCR then denatures the library fragment leading two separate strands, one of which (the reverse strand) anneals to the bead. The annealed DNA is amplified by polymerase starting from the bead towards the primer site. The original reverse strand then denatures and is released from the bead only to re-anneal to the bead to give two separate strands. These are both amplified to give two DNA strands attached to the bead. The process is then repeated over 30-60 cycles leading to clusters of DNA. This technique has been criticised for its time consuming nature, since it requires many steps (forming and breaking the emulsion, PCR amplification, enrichment etc) despite its extensive use in many of the NGS platforms. It is also relatively inefficient since only around two thirds of the emulsion micro reactors will actually contain one bead. Therefore an extra step is required to separate empty systems leading to more potential inaccuracies.

Bridge PCR

The surface of the flow cell is densely coated with primers that are complementary to the primers attached to the DNA library fragments (Figure 3). The DNA is then attached to the surface of the cell at random where it is exposed to reagents for polymerase based extension. On addition of nucleotides and enzymes, the free ends of the single strands of DNA attach themselves to the surface of the cell via complementary primers, creating bridged structures. Enzymes then interact with the bridges to make them double stranded, so that when the denaturation occurs, two single stranded DNA fragments are attached to the surface in close proximity. Repetition of this process leads to clonal clusters of localised identical strands. In order to optimise cluster density, concentrations of reagents must be monitored very closely to avoid overcrowding.

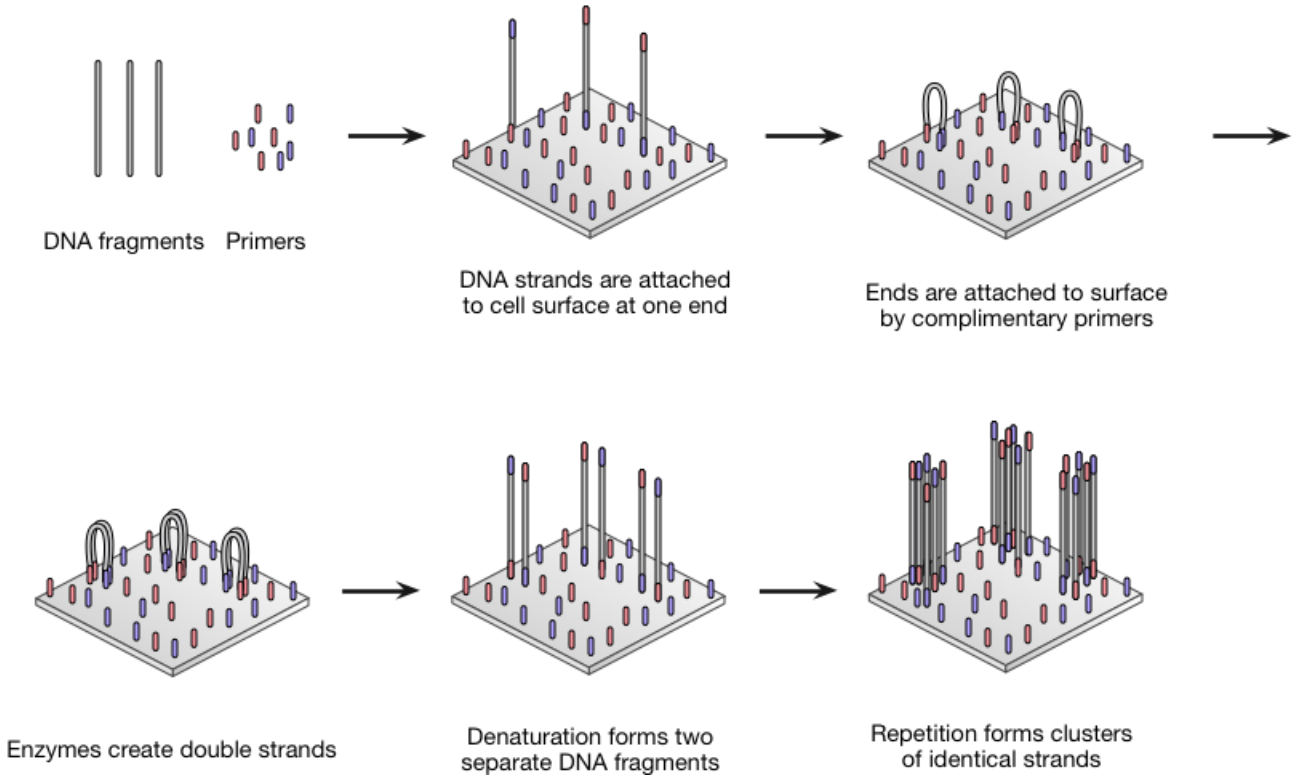


Figure 3 | Bridging PCR

SEQUENCING

Several competing methods of Next Generation Sequencing have been developed by different companies.

454 Pyrosequencing

Pyrosequencing is based on the 'sequencing by synthesis' principle, where a complementary strand is synthesised in the presence of polymerase enzyme (Figure 4). In contrast to using dideoxynucleotides to terminate chain amplification (as in Sanger sequencing), pyrosequencing instead detects the release of pyrophosphate when nucleotides are added to the DNA chain. It initially uses the emulsion PCR technique to construct the polonies required for sequencing and removes the complementary strand. Next, a ssDNA sequencing primer hybridizes to the end of the strand (primer-binding region), then the four different dNTPs are then sequentially made to flow in and out of the wells over the polonies. When the correct dNTP is enzymatically incorporated into the strand, it causes release of pyrophosphate. In the presence of ATP sulfurylase and adenosine, the pyrophosphate is converted into ATP. This ATP molecule is used for luciferase-catalysed conversion of luciferin to oxyluciferin, which produces light that can be detected with a camera. The relative intensity of light is proportional to the amount of base added (i.e. a peak of twice the intensity indicates two identical bases have been added in succession).

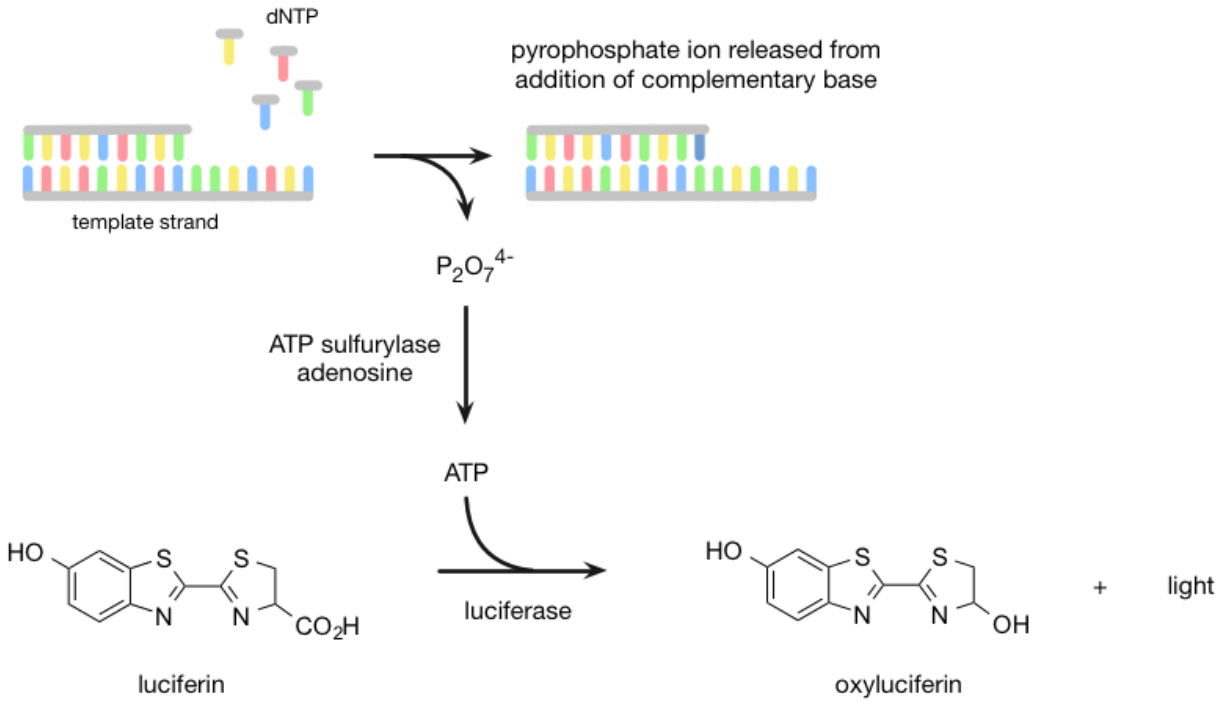


Figure 4 | 454 Pyrosequencing

Pyrosequencing, developed by 454 Life Sciences, was one of the early successes of Next-generation sequencing; indeed, 454 Life Sciences produced the first commercially available Next-generation sequencer. However, the method was eclipsed by other technologies and, in 2013, new owners Roche announced the closure of 454 Life Sciences and the discontinuation of the 454 pyrosequencing platform.

Ion torrent semiconductor sequencing

Ion torrent sequencing uses a "sequencing by synthesis" approach, in which a new DNA strand, complementary to the target strand, is synthesized one base at a time. A semiconductor chip detects the hydrogen ions produced during DNA polymerization (Figure 5).

Following polony formation using emulsion PCR, the DNA library fragment is flooded sequentially with each nucleoside triphosphate (dNTP), as in pyrosequencing. The dNTP is then incorporated into the new strand if complementary to the nucleotide on the target strand. Each time a nucleotide is successfully added, a hydrogen ion is released, and it detected by the sequencer's pH sensor. As in the pyrosequencing method, if more than one of the same nucleotide is added, the change in pH/signal intensity is correspondingly larger.

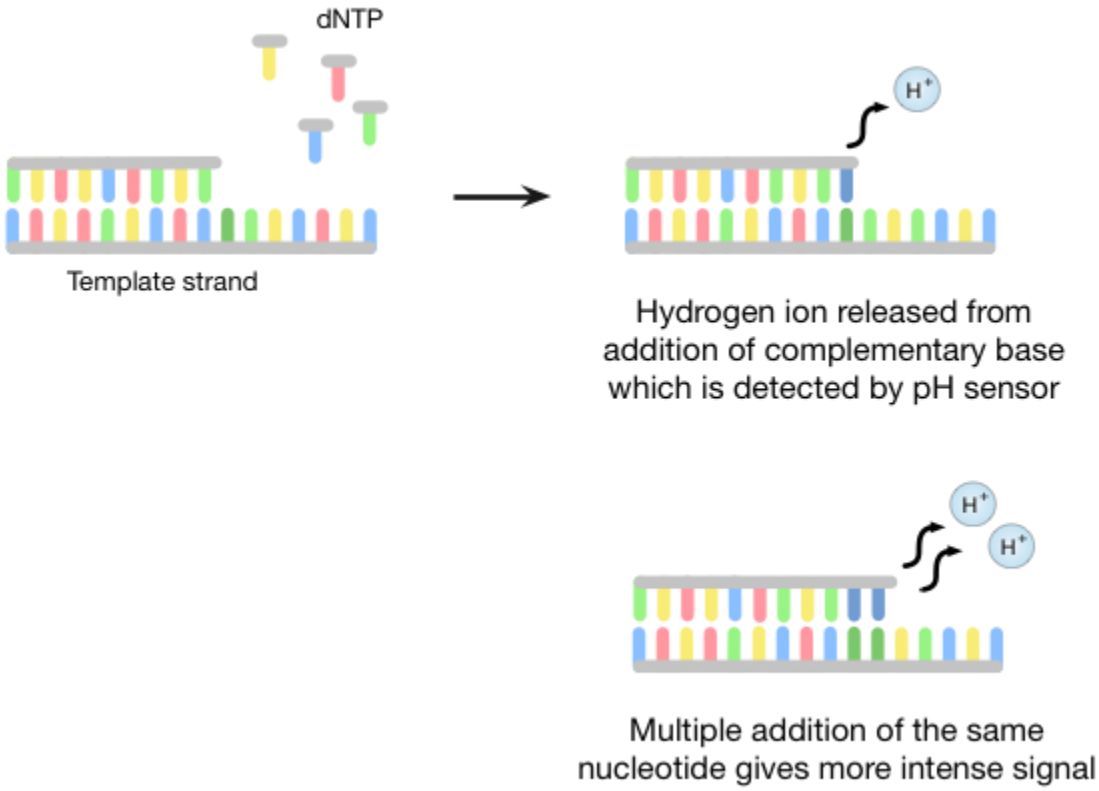


Figure 5 | Ion Torrent semiconductor sequencing

Ion torrent sequencing is the first commercial technique not to use fluorescence and camera scanning; it is therefore faster and cheaper than many of the other methods. Unfortunately, it can be difficult to enumerate the number of identical bases added consecutively. For example, it may be difficult to differentiate the pH change for a homorepeat of length 9 to one of length 10, making it difficult to decode repetitive sequences.

Sequencing by ligation (SOLiD)

SOLiD is an enzymatic method of sequencing that uses DNA ligase, an enzyme used widely in biotechnology for its ability to ligate double-stranded DNA strands (Figure 6). Emulsion PCR is used to immobilise/amplify a ssDNA primer-binding region (known as an adapter) which has been conjugated to the target sequence (i.e. the sequence that is to be sequenced) on a bead. These beads are then deposited onto a glass surface − a high density of beads can be achieved which which in turn, increases the throughput of the technique.

Once bead deposition has occurred, a primer of length N is hybridized to the adapter, then the beads are exposed to a library of 8-mer probes which have different fluorescent dye at the 5' end and a hydroxyl group at the 3' end. Bases 1 and 2 are complementary to the nucleotides to be sequenced whilst bases 3-5 are degenerate and bases 6-8 are inosine bases. Only a complementary probe will hybridize to the target sequence, adjacent to the primer. DNA ligase is then uses to join the 8-mer probe to the primer. A phosphorothioate linkage between bases 5 and 6 allows the fluorescent dye to be cleaved from the fragment using silver ions. This cleavage allows fluorescence to be measured (four different fluorescent dyes are used, all of which have different emission spectra) and also generates a 5’-phosphate group which

can undergo further ligation. Once the first round of sequencing is completed, the extension product is melted off and then a second round of sequencing is perfomed with a primer of length N−1. Many rounds of sequencing using shorter primers each time (i.e. N−2, N−3 etc) and measuring the fluorescence ensures that the target is sequenced.

Due to the two-base sequencing method (since each base is effectively sequenced twice), the SOLiD technique is highly accurate (at 99.999% with a sixth primer, it is the most accurate of the second generation platforms) and also inexpensive. It can complete a single run in 7 days and in that time can produce 30 Gb of data. Unfortunately, its main disadvantage is that read lengths are short, making it unsuitable for many applications.

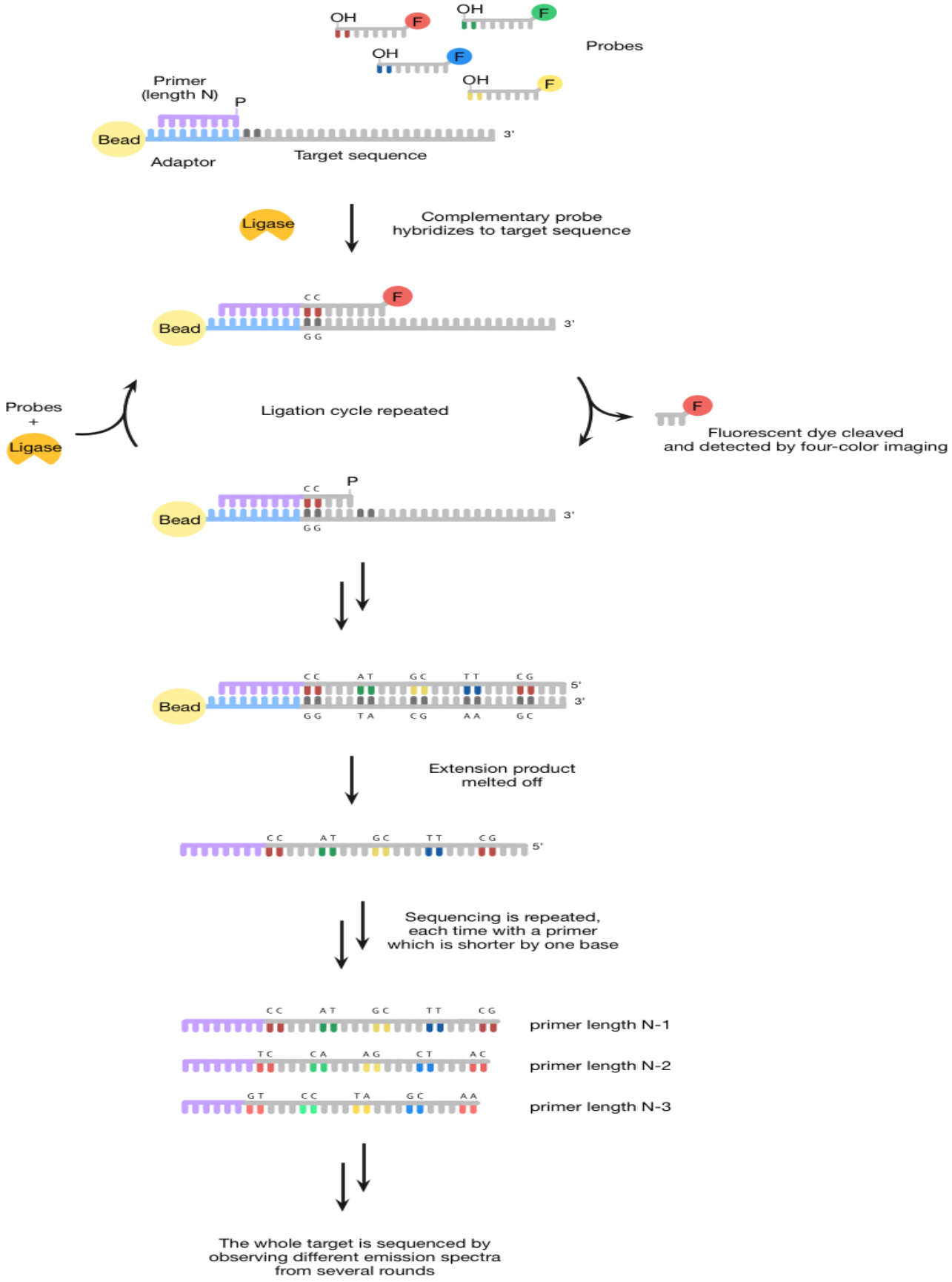


Figure 6 | Sequencing by ligation

Reversible terminator sequencing (Illumina)

Reversible terminator sequencing differs from the traditional Sanger method in that, instead of terminating the primer extension irreversibly using dideoxynucleotide, modified nucleotides are used in reversible termination. Whilst many other techniques use emulsion PCR to amplify the DNA library fragments, reversible termination uses bridge PCR, improving the efficiency of this stage of the process.

Reversible terminators can be grouped into two categories: 3′-O-blocked reversible terminators and 3′-unblocked reversible terminators.

3′-O-blocked reversible terminators

The mechanism uses a sequencing by synthesis approach, elongating the primer in a stepwise manner. Firstly, the sequencing primers and templates are fixed to a solid support. The support is exposed to each of the four DNA bases, which have a different fluorophore attached (to the nitrogenous base) in addition to a 3'-O-azidomethyl group (Figure 7).

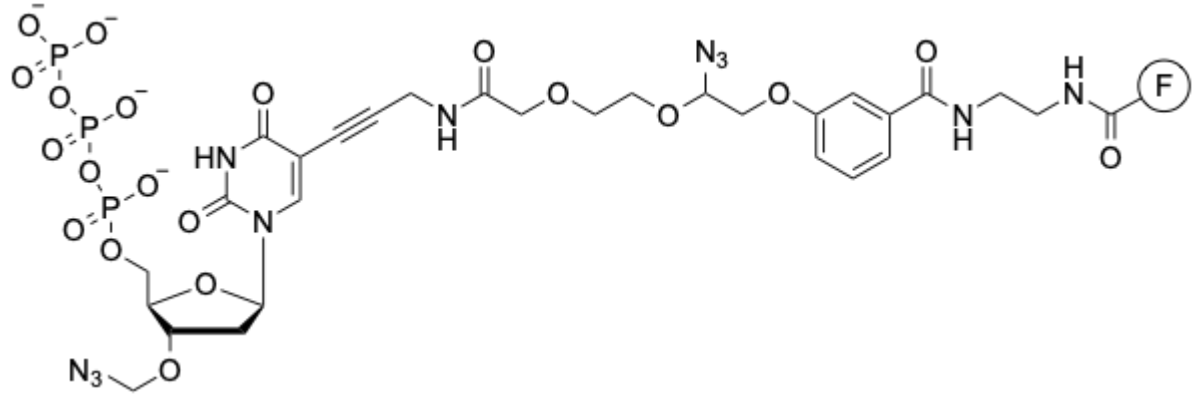


Figure 7 | Structure of fluorescently labelled azidomethyl dNTP used in Illumina sequencing

Only the correct base anneals to the target and is subsequently ligated to the primer. The solid support is then imaged and nucleotides that have not been incorporated are washed away and the fluorescent branch is cleaved using TCEP (tris(2-carboxyethyl)phosphine). TCEP also removes the 3'-O-azidomethyl group, regenerating 3'-OH, and the cycle can be repeated (Figure 8) .

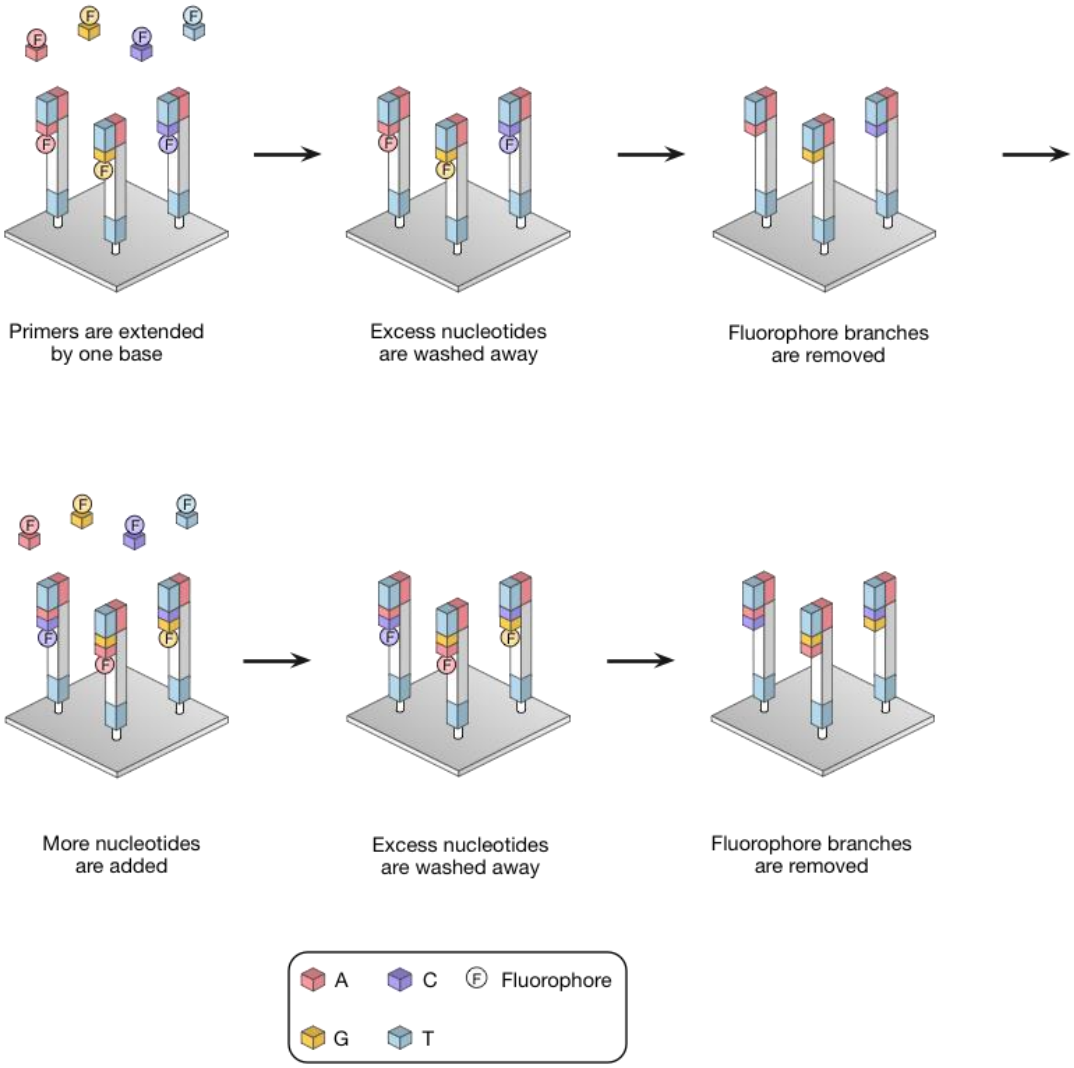


Figure 8 | Reversible terminator sequencing

3′-unblocked reversible terminators

The reversible termination group of 3′-unblocked reversible terminators is linked to both the base and the fluorescence group, which now acts as part of the termination group as well as a reporter. This method differs from the 3′-O-blocked reversible terminators method in three ways: firstly, the 3’-position is not blocked (i.e. the base has free 3’-OH); the fluorophore is the same for all four bases; and each modified base is flowed in sequentially rather than at the same time.

The main disadvantage of these techniques lies with their poor read length, which can be caused by one of two phenomena. In order to prevent incorporation of two nucleotides in a single step, a block is put in place, however in the event of no block addition due to a poor synthesis, strands can become out of phase creating noise which limits read length. Noise can also be created if the fluorophore is unsuccessfully attached or removed. These problems are prevalent in other sequencing methods and are the main limiting factors to read length.

This technique was pioneered by Illumina, with their HiSeq and MiSeq platforms. HiSeq is the cheapest of the second generation sequencers with a cost of $0.02 per million bases. It also has a high data output of 600 Gb per run which takes around 8 days to complete.

NGS Technologies Overview

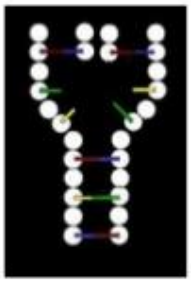
- NGS differs in template preparation, sequencing and imaging, and data analysis
- **Commercially available technologies:**
  - Illumina/Solexa
  - Roche/454
  - Helicos BioSciences
  - Life/APG – SOLiD system
  - Pacific Biosciences
  - Ion Torrent technology

Comparison Of NGS Platforms

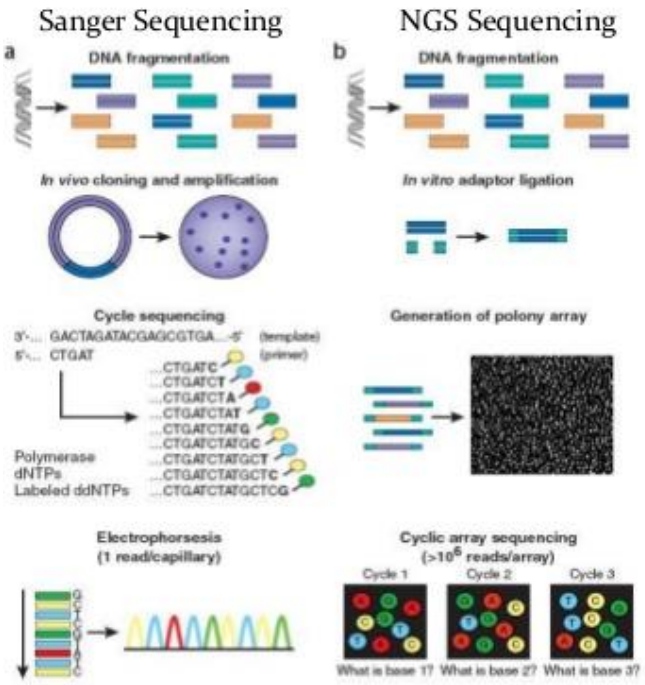
Method	Single-molecule real time sequencing	Ion semiconductor	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	2900 bp average[1]	200 bp	700 bp	50 to 250 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	87% (read length mode), 99% (accuracy mode)	98%	99.9%	98%	99.9%	99.9%
Reads per run	35–75 thousand	up to 5 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases	$2	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Longest read length. Fast. Detects 4mC, 5mC, 6mA.	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Low yield at high accuracy. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive.	Slower than other methods.	More expensive and impractical for larger sequencing projects.

SANGERS Vs. NGS

Features	Sanger	NGS
Sequencing Samples	Clones, PCR	DNA Libraries
Preparation Steps	Few, Sequencing reactions clean up	Many, Complex procedures
Data Collection	Samples in plates : 96, 384	Samples on slides 1-16+
Data	1 Read/ Sample	Thousands & Millions of Reads/ Samples.

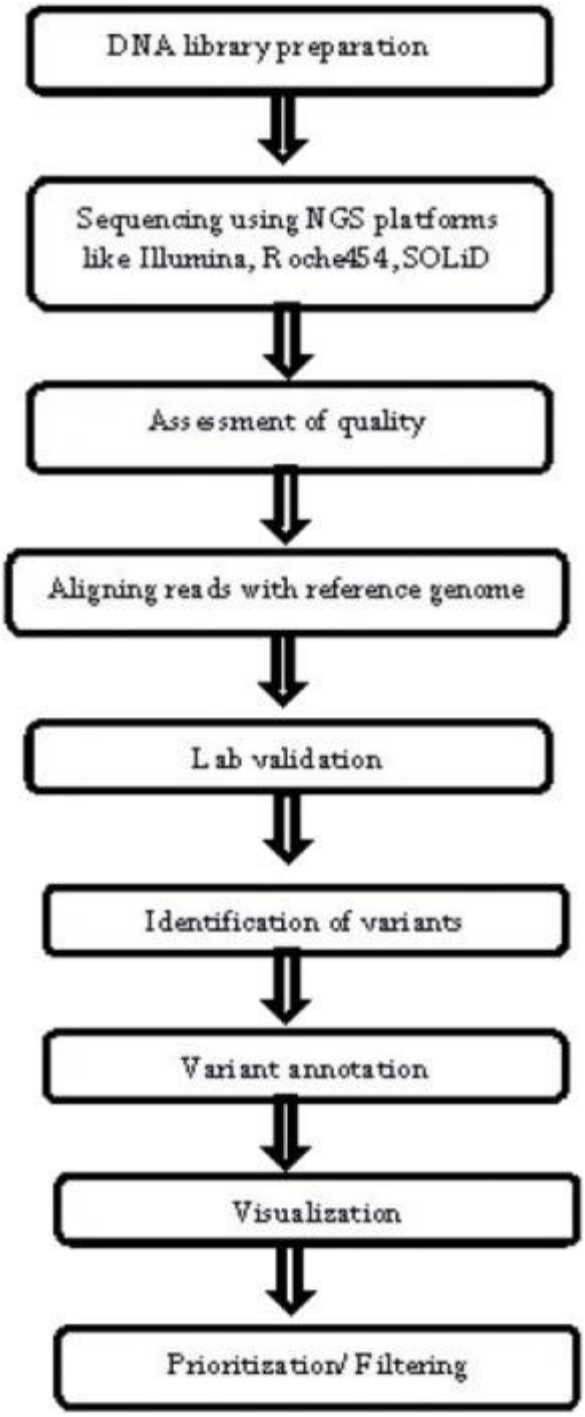


ADVANTAGES OF NGS



- No *invivo* cloning, Transformation, Colony picking
- High degree of Parallelism then Capillary Sequencing
- Low Reagent Cost
- Reduced Sample Size
- Less Time

NGS DATA PROCESSING - Workflow of NGS data analysis



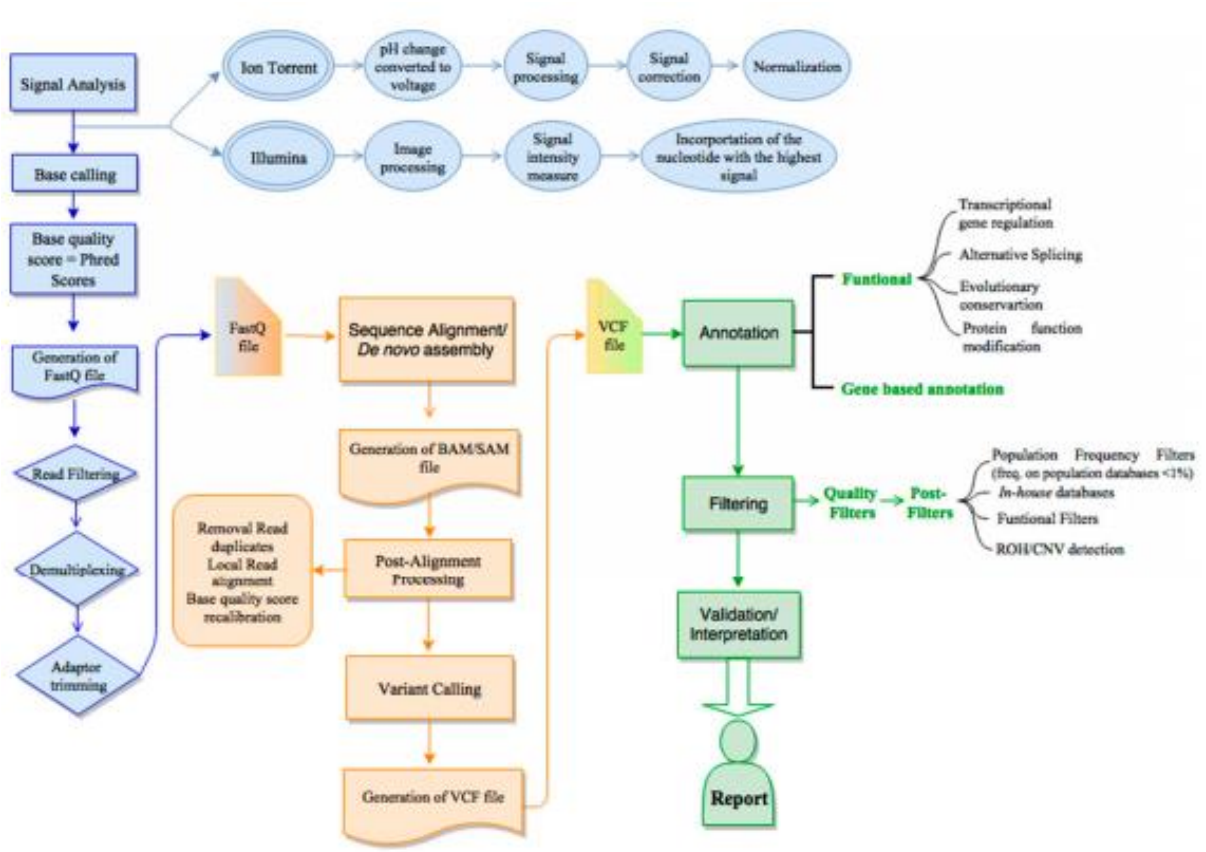


Figure 2. An overview of the next generation sequencing (NGS) bioinformatics workflow. The NGS bioinformatics is subdivided in the primary (blue), secondary (orange) and tertiary (green) analysis. The primary data analysis consists of the detection and analysis of raw data. Then, on the secondary analysis, the reads are aligned against the reference human genome (or *de novo* assembled) and the calling is performed. The last step is the tertiary analysis, which includes the variant annotation, variant filtering, prioritization, data visualization and reporting. CNV—copy number variation; ROH—runs of homozygosity, VCF—variant calling format.

First, the DNA library is prepared and samples are sequenced using NGS platform. Then, quality assessment of NGS reads is carried out and reads are aligned with the reference genome. After that, variant identification and annotation is performed followed by visualization. Further prioritization and filtration of identified variations is followed by validation of the generated results in the lab (Fig. 4). NGS instruments give higher throughput data at an immense speed by sequencing a huge number of short DNA fragments in parallel [56,57]. The three most commonly utilized platforms Roche 454, Illumina and ABI SOLiD sequence DNA by measuring and analyzing signals, which are discharged amid the formation of the second DNA strand, however the contrast in how the second strand is created. Keeping in mind the end goal to create detectable signals, template DNA is divided into small fragments, amplified and immobilized on a glass slide before sequencing. Subsequent to finishing lab work and the real sequencing, the researcher will have a huge amount of raw data to be further processed. The analysis of the data can be divided into five particular steps (Fig. 4): i) quality assessment of the raw data, (ii) read alignment to a reference genome, (iii) variant identification, (iv) annotation of the variants and (v) data visualization

Assessment of quality

In this step, quality of NGS reads is evaluated to remove, correct or trim the reads not meeting the standards. Errors such as base calling errors, poor quality reads etc. are assessed in this step [58]. For this, tools such as FASTQC are used, which assesses the quality by considering the above mentioned errors with calculation of quality scores.

Aligning sequences

After assessing the quality of NGS reads, the reads are aligned to the reference genome. For that UCSC (University of Santa Cruz) and GRC (Genome Reference Consortium) are mainly used as sources of human reference genome [59–61]. There are some issues in selecting alignment software, the first is solving the problem of ambiguity in mapping short reads to the reference genome, which can be solved by considering paired-end reads as a better option [62]. Secondly, mutations generated from reads with many mismatches have to be discarded from further analysis steps.

Identifying variants

Variant identification is a very important part of NGS data analysis. In this, sequence coverage is a main parameter, as identified mutations should be supported by several reads [63]. Tools of variant identification are divided into 4 categories: (i) germline callers, (ii) somatic callers, (iii) Copy Number Variants (CNV) identification and (iv) Structural Variants (SV) identification. In case of rare diseases, germline mutations are focused while in cancer, somatic mutations are targeted for detection. Structural variant identification tools identify SVs such as inversions, translocations or large INDELS as well as CNVs which are the simplest form of SVs only [64]. The list of variant identification tools is provided in Table 2.

Annotating variants

Annotation of variants provides biological significance by identifying disease causing variants. Annotation of SNPs and INDELs is provided via computational annotation tools by providing links to pertaining databases such as dbSNP etc. Yohe S and co-workers have identified and annotated 1102 variants of inherited disorders across the 568 genes using the ANNOVAR tool combined with SIFT, PolyPhen2 and Provean annotation scores to evaluate the functional significance of novel variants [65]. The various variant annotation tools are listed in Table 3.

Visualizing NGS data

After annotating the variants, they are visualized using visualization tools and genome browsers. By visualizing the variants, we can obtain information about variants, such as mapping quality, aligned reads, annotation information which includes consequence, impact of variants, scores of different annotation tools, etc. [66]. Paula Paulo and coworkers have identified and visualized functionally deleterious germline mutations in novel genes in early-onset/familial prostate cancer using Geneticist Assistant tool [67].

NGS tools selection criteria

For identifying variations, germline and somatic variant callers can be selected if: (i) they use Binary Alignment/Map (BAM) or pileup and Sequence alignment/map (SAM) [68] format as input and (ii) the tools offer output effects in the Variant Call Format (VCF). SVs and CNVs detection tools are used after the acceptance of SAM/BAM as input format. For annotating variations, requirements of annotation packages are: (i) It should accept VCF as input format and (ii) It should integrate results from other software. GUI availability, VCF, SAM and BAM support are required for visualization of results.

Table 2
Tools for variant identification.

Tools	Input files	Output files	Identifies
Germline caller tools			
Galaxy platform	BAM/SAM	VCF, VarScan CSV	SNP, INDEL
SanGeniX platform*	BAM/SAM	VCF, VarScan CSV	SNP, INDEL
VarScan2	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL
SNVer	BAM/SAM	VCF	SNP, INDEL
CRISP	BAM/SAM	VCF	SNP, INDEL
GATK(Unified Genotyper)	BAM/SAM	VCF	SNP, INDEL
SAMtools	BAM/SAM, FASTA	VCF	SNP, INDEL
Somatic callers tools			
Galaxy platform	BAM/SAM	VCF, VarScan CSV	SNP, INDEL
SanGeniX platform*	BAM/SAM	VCF, VarScan CSV	SNP, INDEL
VarScan2	pileup/mpileup	VCF, VarScan CSV	INDEL, SNP,CNV
GATK (Somatic Indel Detector)	BAM/SAM	VCF	INDEL
SAM tools	BAM/SAM, FASTA	BCF	SNP, INDEL
CNV identification tools			
ExomeCNV	BAM/SAM, pileup + bed + FASTA	CSV	CNV, LOH
CNVnator	BAM/SAM, FASTA	CSV	CNV
CONTRA	BAM/SAM, FASTA	VCF, CSV	CNV
RDXplorer	BAM/SAM, FASTA	CSV	CNV
SV identification tools			
GASVPro (GASVPro-HQ)	BAM/SAM	clusters file	INDEL, INV, TRANS
CLEVER	BAM/SAM, FASTA	CLEVER format	INDEL
BreakDancer	BAM/SAM, config file	BED, CSV	INDEL, CNV, INV, TRANS
Breakpointer	BAM/SAM	GFF	INDEL
SVMerge	BAM/SAM, FASTA	BED	INDEL, INV, CNV

*Proprietary software.

Table 3
Tools for variant annotation.

Tools	Input files	Output files	Variants
Galaxy platform	VCF, pileup/TXT	VCF, TXT, HTML overview	SNP, INDEL
SanGeniX platform*	VCF, pileup/TXT	VCF, TXT, HTML overview	SNP, INDEL
VARIANT	VCF,BED, GFF2	TXT, web report,	SNP, INDEL, CLI, Web
snpEff	VCF	VCF, HTML overview TXT	SNP, INDEL, CLI
NGS -SNP	VCF, pileup, MAQ, diBayes, TXT	TXT	SNP,CLI
VEP	VCF, pileup, HGVS, TXT, variant Identifiers	TXT	SNP, INDEL,CLI, Web
ANNOVA-R	VCF, GFF3-SOLiD, SOAPsnp, pileup, CompleteGenomics, MAQ, CASAVA	TXT	SNP, INDEL, CNV, CLI,
AnnTools	VCF, pileup, TXT	VCF	SNP, INDEL, CNV, CLI
SeattleSeq	VCF, GATK BED, custom, MAQ, CASAVA	VCF, SeattleSeq	SNP, INDEL, Web
SVA	VCF, SV.events file, BCO	CSV	SNP, INDEL, CNV, GUI, CLI

*Proprietary software.

Structural variome

Variation in more than one nucleotide is called the structural variome. There are two major classes of structural variations: balanced and unbalanced variations. Balanced variations do not change content of DNA while unbalanced variations change the content of DNA. Inversion, same chromosomal translocation and different chromosomal translocation are subtypes of balanced structural variations, while duplication and deletion are subtypes of unbalanced structural variations. The types of known structural variations are highlighted in Fig. 5. Structural variations can be detected by five types of methods. First is the Pair-end mapping (PEM) method. In this type, the two ends of the DNA fragment are sequenced and uniquely mapped to the reference genome. Second is the Single-end method in which single ends of multiple DNA fragments are sequenced and mapped with the reference genome at

different positions, which forms overlapping in read mapping. The third type is Translocation and inversion detection. In interchromosomal translocations, one member of the pair maps to one chromosome and its mate to every other. And in inversions or intrachromosomal translocations, the two ends map to the equal chromosome, however in the wrong orientation or the wrong distance apart. Fourthly is Copy number variant detection which can be defined as stretches of DNA, longer than a kilobase, which is present in the genome with an abnormal number of copies that include large deletions and duplications, as well as unbalanced translocations. Large deletions are less difficult to detect than smaller indels using paired-end methods, as they are easily identified from normal variation within the insert size. Large duplications are harder to discover, as there may be no single read or read pair spanning the insertion. And the fifth type is Insertion and deletion detection. Indels are common in human genome and make a contribution to genetic diversity and human diseases [71–73]. In the clinical molecular oncology laboratory, the detection of small (< 10 bp) and medium (> 10 but < 1 kb) indels is important to many cancers. Of specific clinical importance are the NPM1 insertion, FLT3 internal tandem duplication (FLT3-ITD), KIT exon 8 indels in acute myeloid leukemia and EGFR exons 19 and 21 insertions and deletions in lung cancers [74–77]. By Sanger sequencing or gel capillary based sizing methods, small-and medium sized indels are typically simple to detect. Indel detection via NGS methods has been challenging largely because of the short read lengths generated by using NGS methods. In general, small indels can be called with reasonable sensitivity from NGS data, despite the fact that the specificity has a tendency to be low. Further, most indel detection software detect deletions over insertions because of inherited bias in the tool, as inserted sequences are more difficult to align to the reference sequences.

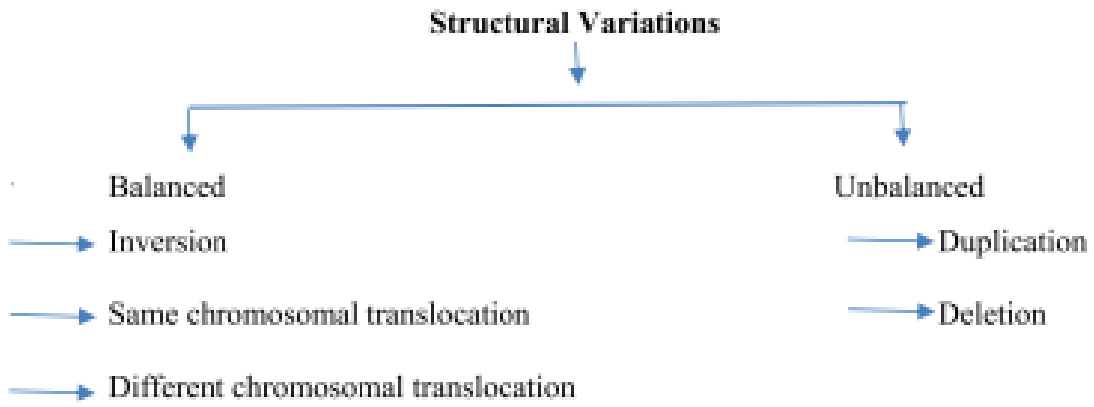


Fig. 5. Structural variations.

Clinical validation of variants

Translation of NGS methods for clinical use is a very important and challenging task that is carried out by validation of different performance characteristics. Clinical validation of NGS data is performed by measuring different parameters like analytical sensitivity that is defined

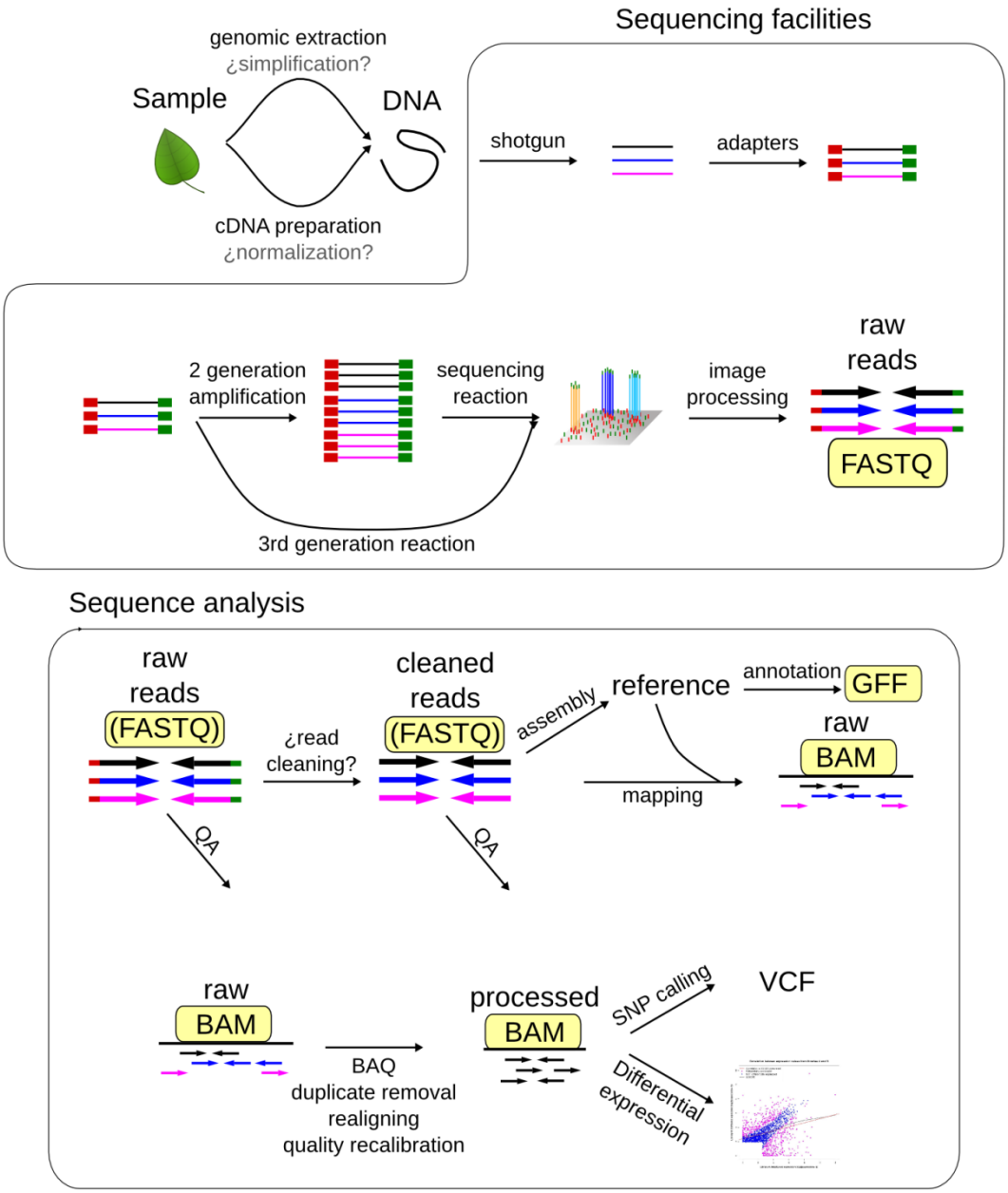
as an ability of the assay to detect true sequence variants i.e. falsenegative rate, analytical specificity that is defined as probability of the assay to not detect mutations where none are present i.e. false-positive rate. Accuracy is the measure of sequencing accuracy and error rates and precision are the measure of reproducibility of mutation detection by the assay and inter-user reproducibility. In a related work, researchers have clinically validated 30 known mutations of more than 100 inherited diseases with 100% analytical sensitivity and 100% analytical specificity for 18 samples of patients in whom pathogenic mutations were previously identified

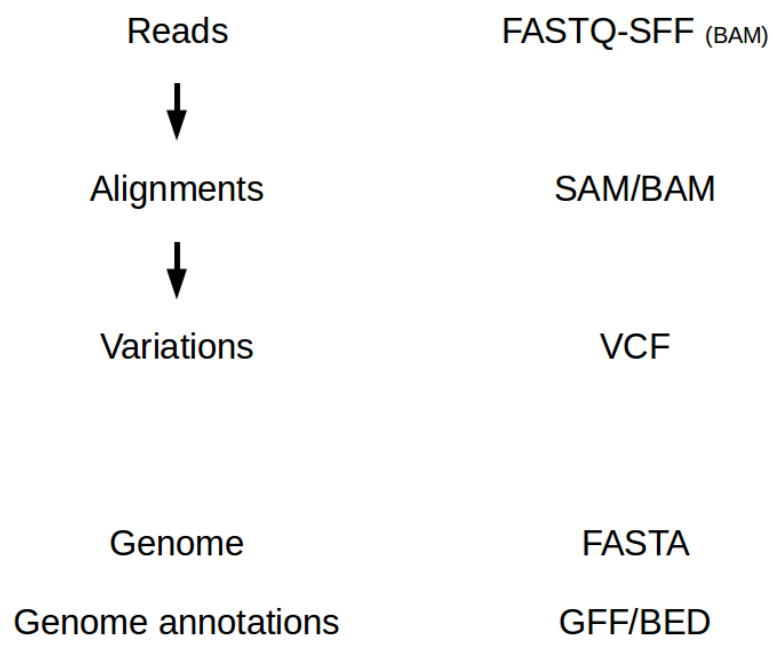
Handling of NGS data

While analyzing NGS data, a number of intermediate analysis files and result files are generated that are collectively very large in size i.e., 100's of GB, TB and even reaching petabytes. Interpretation of these complex NGS data files especially for aggregated large amounts of variations or heterogeneous sequencing data, is challenging in terms of translating data to knowledge for clinical applications. Also, processing power, memory (RAM) and data storage are hardware bottlenecks in computational analysis that can be overcome by high performance computational resources, but increase the computational cost [84]. After analyzing the NGS data, the next step is handling of the resultant NGS data, which is carried out by employing machine learning based methods. Machine learning is a descendent of the statistical model fitting method, which extracts the information from data by building probabilistic models. Efficient machine learning methods study huge amounts of generated NGS data comparatively and evolutionarily [85]. Good classification and regression results can be yielded by machine learning methods like Support Vector Machines, Artificial Neural Network. Weizhong Zhao and coworkers have employed machine learning methods for evaluating the performance of the NGS data set for the Salmonella enterica strains [86]. The analysed NGS data can be classified with these methods to obtain clinically significant results.

NGS FILE FORMATS

There are lots of file formats related to NGS analyses. The most common ones are:





Sequence file formats

The different sequence related formats include different information about the sequence. The most common file formats in the NGS world are: fastq and sff.

SFF

The SFF (Standard Flowgram Format) files are the 454 equivalent to the ABI chromatogram files. They hold information about:

- the flowgram,
- the called sequence,
- the qualities,
- and the recommended quality and adaptor clipping.

These recommended clippings are given by the 454 sequencer. The Roche software takes into account the quality and the adaptor sequence to recommend a clipping for each sequence. Like the ABI files, these are binary files that should be opened with specialized programs. There are several tools to extract the sequences and to convert them to a more usable format. Roche provides one executable able to do it with the 454 machine. Alternatively we can use the sff_extract tool to obtain a fasta file.

Fasta

The fasta format is based on a simple text. Each sequence starts with a “>” followed by the sequence name, an space and, optionally, the description

```
>seq_1 description
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCAC
AGTTT
>seq_2
ATCGTAGTCTAGTCTATGCTAGTGCGATGCTAGTGCTAGTCGTATGCATGGCTAT
GTGTG
```

Usually, if we have quality information, another fasta file with the quality information could be provided. In this cases both the sequence and the quality file should have the sequences in the same order.

sanger fastq

The fastq format was developed to provide a convenient way of storing the sequence and the quality scores in the same file. These are text files and they look like:

```
@seq_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCAC
AGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65
@seq_2
ATCGTAGTCTAGTCTATGCTAGTGCGATGCTAGTGCTAGTCGTATGCATGGCTAT
GTGTG
+
208DA8308AD8SF83FH0SD8F08APFIDJFN34JW830UDS8UFDSADPFIJ3N8DAA
```

In this file every sequence has 4 lines. In the first line we get the sequence name after the symbol “@” and, optionally, the description. The second line has the sequence and the fourth line has the quality scores encoded as letters.

Illummina fastq

This file is almost identical to a sanger fastq file, but the encoding for the quality scores is different. When we deal with a fastq file we have to be sure about which kind of file we are dealing with, an illumina fastq or a sanger fastq. Unfortunately they are not easy to differentiate. Also you have to take into account that solexa used to had a third fastq format, the solexa fastq, although this one is mostly obsoleted. Recently Illumina has also decided to distribute its files as Sanger fastq, so the Illumina fastq will be not used any more.

One of the seq_crumbs utilities, guess_seq_format, is able to differentiate the Sanger from the Illumina version by looking for quality characters exclusive of the Sanger version.

SRA

SRA is the file format in which all NCBI SRA content is provided. SRA files are binary files and we need specific tools to extract the information. There is a toolkit (SRA Toolkit)developed by NCBI to deal with these binary files.

Compressed files

Sometime these sequence text file can be found compressed to save up hard drive space. The most common compression formats are gzip and bgzip. bgzip is a gzip variant commonly used in genomics because, although it is a little less efficient in the compression ratio, it allows random access. Most software is becoming compatible with these formats.

Paired files

It is common to obtain two reads from a single molecule. Examples of these techniques are the Illumina pair-ends and mate-pairs. In this cases for each read there is another paired read. One common way to store those paired reads is to create to fastq files, one for the first read of the pairs and another one for the second. In this case the files should hold the reads exactly in the same order.

Fastq file 1
@molecule_1 1st_read_from_pair
@molecule_2 1st_read_from_pair
@molecule_3 1st_read_from_pair

Fastq file 2
@molecule_1 2nd_read_from_pair
@molecule_2 2nd_read_from_pair
@molecule_3 2nd_read_from_pair

Another option is to interleave the reads in a single file alternating the first and second read for each pair.

Interleaved Fastq file
@molecule_1 1st_read_from_pair
@molecule_1 2nd_read_from_pair
@molecule_2 1st_read_from_pair
@molecule_2 2nd_read_from_pair
@molecule_3 1st_read_from_pair
@molecule_3 2nd_read_from_pair

Depending on the software that we want to use we should the interleaved or the two file version. In seq_crumbs there are programs to convert between one option and the other.

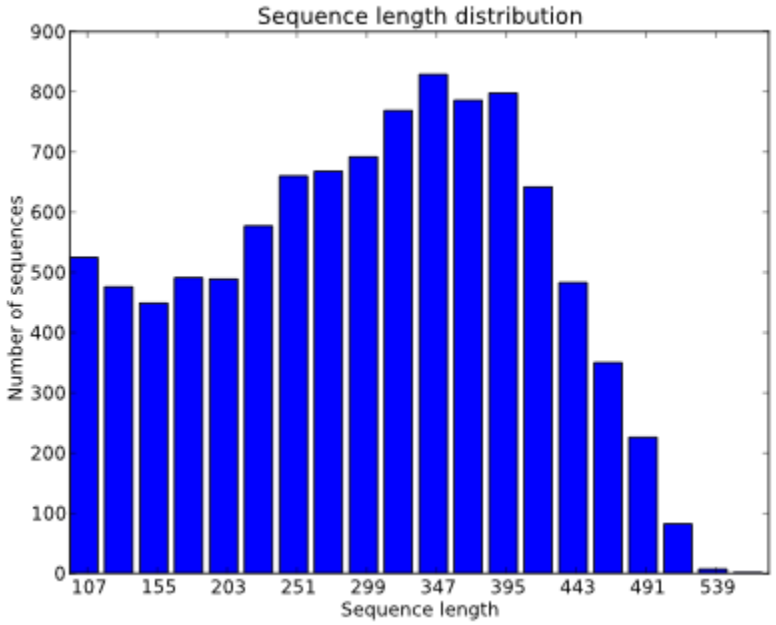
Read QA and Cleaning

Before using the raw sequences generated by the sequencing machines we have to check their quality and eventually to clean them to get rid of adaptor, contaminants and low quality regions.

Assessing the read quality

We can assess the quality of the reads by taking a look at their length distribution, phred quality distribution, nucleotide frequencies and complexity. It is highly recommended to take a look at the excellent documentation found in the fastqc and prinseq sites.

The length distribution of the reads is a basic quality check. We have to make sure that the length distribution complies with the expected distribution for the sequencing technology that we have used.

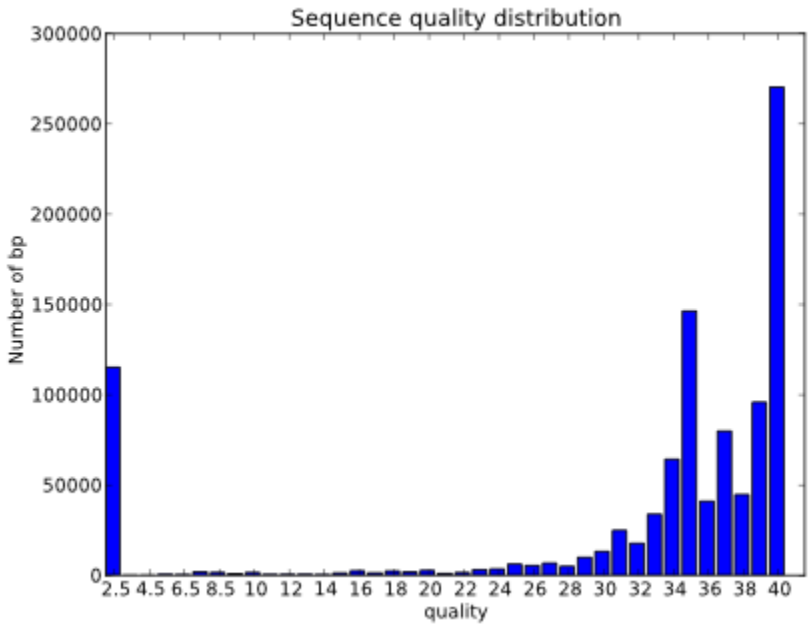


For Illumina it would be typical to obtain the same sequence length for all reads.

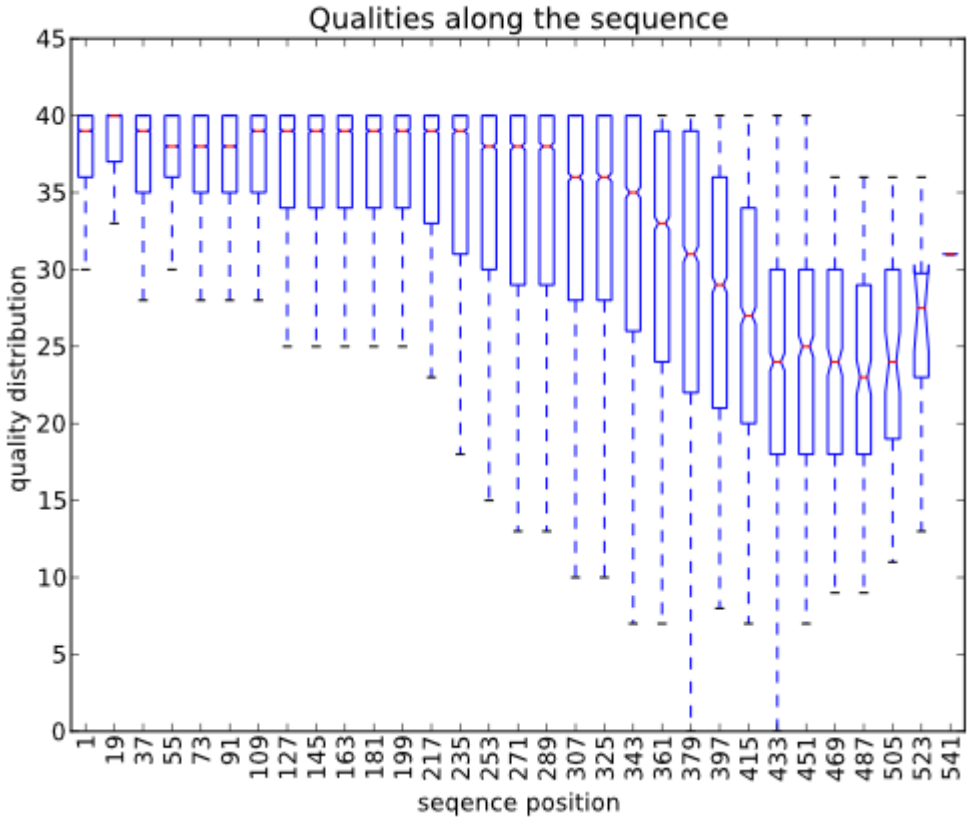
We should also evaluate the sequence quality. We can do it by calculating some statistics like: mean quality, Q20 and Q30. Q30 is the percentage of bases in the reads with a phred quality equal or bigger than 30. For instance:

```
Q20: 86.85
Q30: 82.39
minimum: 2
maximum: 41
average: 32.32
```

We can also evaluate the quality by taking a look at the quality distribution.

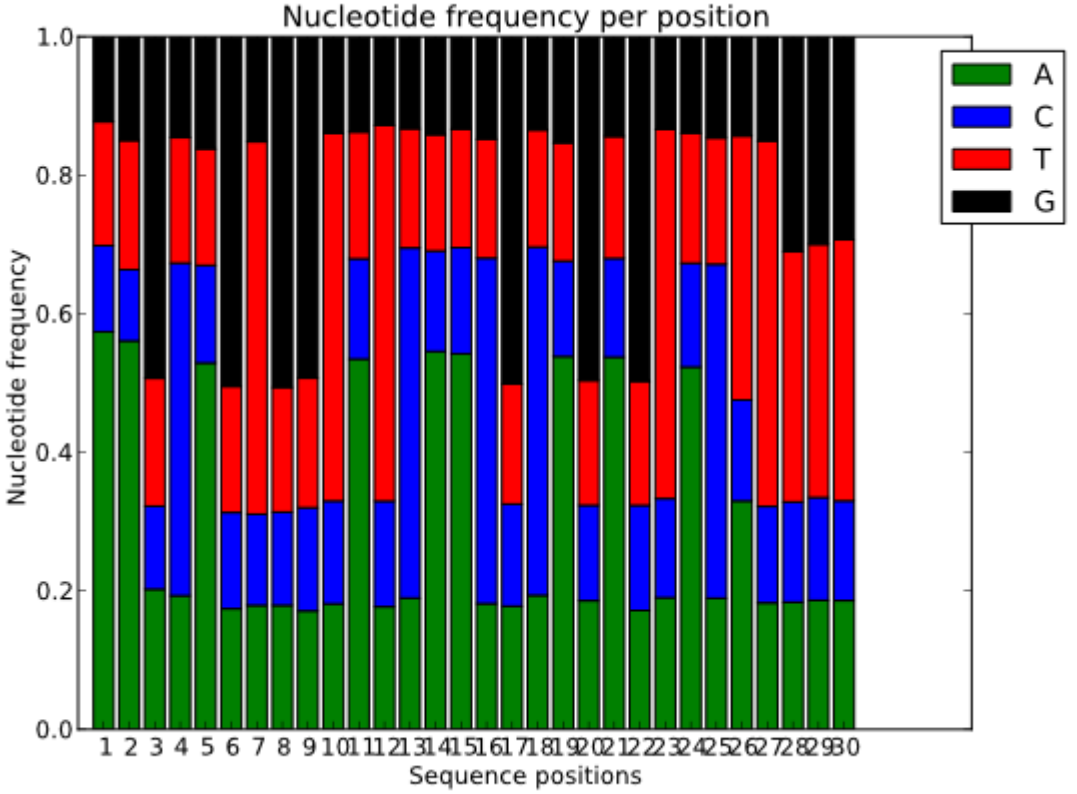


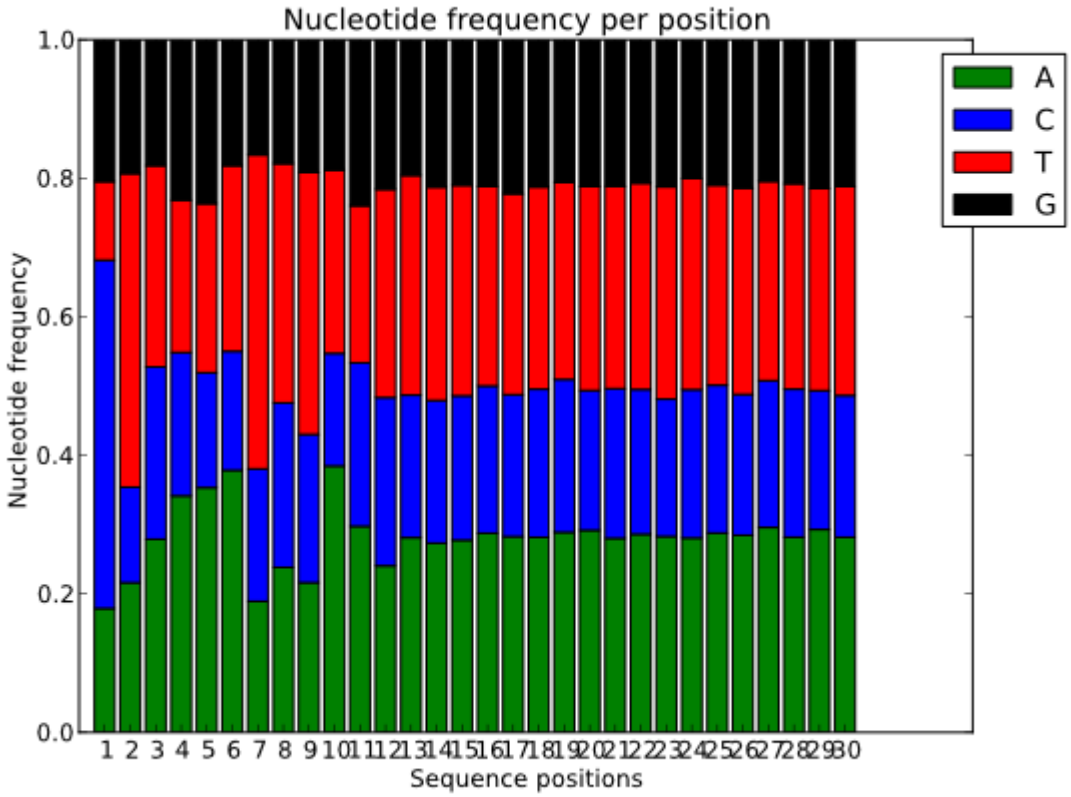
Another very useful and common way of evaluating the quality is to generate a boxplot with the qualities per position along de reads.



Also, to spot the presence of adaptors at the first positions of the reads it is common to represent the frequency of each nucleotide for each position. Ideally, in these charts, all

positions should have the same nucleotide frequencies, but it is common to find adapter contamination or biases produced during the library construction protocol.





It can be also quite useful to study the k-mer composition of the reads. This will give us an idea of the overall complexity of the sequence and also it can serve to spot highly repetitive k-mers corresponding to adapters, poli-A or repetitive sequences.

Read cleaning

The raw sequences can have some regions that could be problematic, for instance vector or adapter sequences and that it would be advisable to remove to avoid problems with downstream analyses. Some of these problems are:

- Vectors
- Adapters
- Low quality
- Low complexity
- Contaminants
- Duplicates
- Error correction

It would be OK to keep these regions if the downstream analysis software is prepared to use this noisy information or if at least would not be negatively affected by it, but in a lot of real world scenarios the downstream software will be negatively affected or can even choke if we do not get rid of this extra noise. For instance, if we want to map the reads with a local mapper (e.g. bowtie2) not trimming the reads wouldn't be so problematic, because the mapper would be capable of even mapping reads with adapters in it. But if we plan to use a global mapper we would be better off trimming the adapters and the bad quality regions to be able to map more reads.

The cleaning programs could be classified in filters and trimmers. The filtering software remove the reads that do not meet the criteria and the trimming software clip only certain regions.

Vectors

During the cloning and sequencing processes several vectors and adapters are usually added to the sequences. If we were to use these raw sequences these vectors are likely to interfere with the rest of the analyses, although this is highly dependent on the analysis and the software that will be applied to the reads. If we know for sure that the software that we are using is prepared to deal with these vectors we could go on without the cleaning, but otherwise it is advisable to do it.

It is not common to find cloning vectors in the NGS reads because the cloning step is skip in most of these experiments, but they are common in the Sanger sequences. If we want to remove them there are two main approaches to find them. If we know the exact vector and cloning site sequence we could use lucy. lucy looks carefully for the cloning sites and for the given vectors and recommends how to trim the sequences. If we are not sure about which

sequencing vector was used we could blast our reads against the UniVec database and trim the regions with significant blast matches.

Adapters

The main practical difference, in the context of the sequence analysis, between the vectors and adapters is the sequence length, the adaptors are short sequences and they are common in the NGS reads. For the long vectors we could use Blast, but to look for the adapters, that are short, with the standard Blast algorithm is not the best approach. It is better to use the blast-short algorithm also implemented by the NCBI Blast software.

When the adapters are shorter than 15 base pairs the algorithms used by the aligners might fail. An alternative in these case it to look for exact matches or to use the cutadapt software.

The blast-short algorithm is used by seq_crumbs.

Quality

For some analyses it could be advisable to remove the regions of bad quality. Some people advise against the low quality trimming of the sequences, because even the low quality regions have information in them. But it is common for the downstream software not to deal particularly well with the low quality regions of the reads, in this case it is important to remove these regions. The usual approach to get rid of the low quality regions is to do a window analysis setting a threshold for the quality. prinseq or seq_crumbs can do this cleaning. Alternatively lucy can clean the long reads.

If the reads have no quality we can estimate which regions had a poor sequencing quality by looking at the density of Ns found in the sequence.

Duplicates

In theory, we would like to obtain one read from every molecule (template) of the original library, but this is not always the case. Due to the PCR amplification and to the detection systems we could end up with even thousands of reads of some molecules. This PCR amplification problem is particularly noteworthy in the systems that use emulsion PCR and in the mate-pair Illumina libraries. In Illumina there are also optical duplicates due to a cluster being read twice. These optical duplicates can be detected because they will appear very close in the slide. The number of expected sequence duplicates depends on the depth of the sequencing, the type of library and the sequencing technology used.

If we do not remove these duplicated reads from the analysis we could calculate skewed allele frequencies in a SNP calling experiment, or false expression profiles in a RNA-seq experiment, or we could give false assurance to an assembler. The problem with the duplication filtering is that when the reads are removed we could be removing reads covering the same region, but that come originally from different molecules, for instance from the

different chromosomes of a diploid individual. By removing them we could lose some information.

Ideally, two duplicated reads should had the same sequence and we could look for them just by searching for identical sequences, but due to the sequencing errors they could be not identical but merely very similar. If we have a reference genome a usual method to remove these duplicates is to remove them once we have align them to the reference. If we don't have a reference we could look at least for reads that are identical. The software PRINSEQ has a module to filter identical duplicated reads.

Low complexity

Low complexity reads can impact several downstream analyses. These low complex reads can be a burden specially for the assemblers, so, in some cases, it could be advisable to filter them out. The NCBI Blast distribution includes dust, a program to mask low complexity regions. ngs_crumbs also has a low complexity filtering executable.

Contaminants

In the samples to analyze there can be different kinds of contaminants:

- Due to the sample preparation, e.g. E. coli.
- Mitochondrial and chloroplastic in genomic samples.
- rRNA in transcriptomes
- pathogens in infected samples

This contaminants should be minimized during the sample preparation, for instance extracting the genomic DNA from isolated nuclei, but if we have them in our reads we can filter them out by running blast searchers. We can filter the contaminants with ngs_crumbs_ by doing blast searches.

Error correction

By trimming and purging the reads the mean sequencing quality of the resulting reads can be improved, but some information is lost. An alternative has been developed and implemented in several programs that tries to correct the errors in the reads. The overall idea is based on gathering the reads, or parts of them, that correspond to the same genomic region and to assume that the changes in low frequency should be due to sequencing errors. This method has been commonly used in the SOLiD world and a review have been recently published: A survey of error-correction methods for next-generation sequencing.

Some conclusions can be derived from the mentioned review. The proposed algorithms and methods are quite new and differ in some key points: quality of the result, memory and time consumed, and scalability. Different methods have been derived for 454/Ion Torrent and Illumina due to their different error models. As the review explains "error correction with respect to a specific genomic position can be achieved by laying out all the reads covering the

position, and examining the base in that specific position from all these reads. As errors are infrequent and random, reads that contain an error in a specific position can be corrected using the majority of the reads that have this base correctly."

There are methods based on the study of the k-mers frequencies and based on multiple sequence alignments. The authors conclude that these methods are more mature for the Illumina reads due to the popularity and the abundance of them. For Illumina Reptile, HiTEC and ECHO are generally more accurate and have better scalability than other methods. The drawback is that most of the software tested failed with some datasets. Only four programs: HSHREC, Reptile, SOAPec and Coral—succeeded in generating results for all data sets. For 454 and ion-torrent the authors recommend Coral over HSHREC.

The study carried out in this review was done in bacteria and in Drosophila.

After the mentioned review a new method based on a different algorithm has been proposed and implemented in the software lighter. The authors claim that this algorithm is faster and requires much less memory despite having a comparable accuracy to the other algorithms. I've tried bless with good results.

Caution should be taken when applying these methods to pooled samples or to polyploids.

Usual read cleaning software

There are plenty of software to process the reads, but some that we have used are:

- Prinseq.
- Trimmomatic.
- cutadapt
- pregap4 (only for Sanger)
- lucy (long reads, created for Sanger)

Sequence Assembly

Assembly software

Some useful assembly software:

- Staden (Sanger).
- Celera assembler (genomic).
- SOAPdenovo (Illumina, genomic).
- trinity (Illumina and 454, transcriptome).
- Mira (454 and not many Illumina reads).
- iAssembler (454, transcriptome).
- newbler (454).

Assembly versus mapping

Once we have a collection of reads there are two different kinds of analyses. If we do not have any previous genomic information we would have to assemble the reads into a genome or transcriptome, as we have already seen in the assembly section. Alternatively, if we had genome already available we could map our reads against that genome. Although both analyses could seem to be similar they are very different. To assemble a genome is computationally much costly than to do a mapping. Assembling the human genome was a difficult task, re-sequencing and mapping the reads from a new individual is much more amenable.

The main computational difference is that the typical software used to assemble requires a time that depends on the total reads length squared or the genome length squared (or quite a lot of memory) while the mapping is just lineal with the reads length. For a review take a look at Sense from sequence reads, but the take home message is that the assembly is time and memory consuming while the mapping can be done in standard computer.

Also, it is important to notice that the read length is a critical parameter for the assemblies, but this is not the case for the mapping. We can map short reads with ease and high accuracy in most cases. Palmieri and Schlötterer reviewed this aspect in 2009.

Mapping

The mapping is the process of comparing each one of the reads with the reference genome. We will obtain one alignment, or more, between each read and the genome.

Like for any other bioinformatic task there is a lot of mapping software available. The most commonly used programs are bowtie2 and bwa. These tools differ on the algorithm used, the sensitivity, the memory requirements, the speed, and the sequence length requirements.

Seqanswers keep a comprehensive list of mappers. In Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions the authors review some of these programs.

SAM format

In general all mappers render the result in a common file format, the SAM format. This format is not meant for human consumption, although we can open the text version of the file. There is a growing collection of software created for dealing with these files. We can merge, sort, filter, realign and browse them. Some useful programs are:

- samtools and picard to merge, sort and filter.
- IGV for browsing.
- GATK to recalibrate the quality scores, to realign the sequences and many other algorithms.

Also the most common SNP callers would require a SAM file to work.

We can encounter SAM files in two flavors: SAM and BAM. The BAM is the binary version and the SAM is just the equivalent text file. They hold exactly the same information and we can convert between them with samtools. These files are composed of two parts, a header in which the sequences used as references are named and the alignment section in which the alignments for all reads are shown. The read groups are defined also in the header. A read group is a collection of reads that share some characteristics like:

- Sample. Name of the individual sequenced.
- Library.
- Platform. Technology used (454, Sanger, Illumina, Solid)

SAM realignment

The mapping is done read by read (pairwise instead of multiple alignment), so the alignment obtained could present some artifacts. There are a couple of ways to avoid these artifacts. One is to inspect the alignment in order to realign the regions with problems to fix them. Another is to mark those problematic regions in order to avoid calling SNPs in there.

The GATK software has an option to realign a BAM file generating a new one with these problems solved. It would be specially advisable running this analysis, specially if we are going to take into account the small indels.

```
reference CAATC   realignment CAATC
read1    CA-TC     ---->    CA-TC
read2    C-ATC              CA-TC
```

samtools has the option (calmd) to calculate a probability for each position in the BAM file of having alignment artifacts. samtools calculates a probability for each position of being incorrectly aligned. The results is a Phred-scaled probability called Base Alignment Quality (BAQ). This BAQ can be combined with the sequencing quality to obtain the probability for each position of being a sequening error or a misalignment.

Duplicated reads

The reads that originate from the same original template are considered duplicated, as we already discussed in the read cleaning section. These duplicated reads align exactly at the same position on the reference genome because their sequence starts exactly at the same point.

If we ignored the sequencing errors the duplicated reads should had exactly the same sequence, but there will be errors. One way to detect them is to look for sequences that are almost identical (the differences being to the sequencing errors) and that align exactly in the 5’ end. If we had pair ends both the forward and the reverse sequencing would had to match and they would be detected more easily. This detection of duplicates is eased once we have

all reads mapped to the reference, so in practice unless we're assembling it tends to be carried out on the BAM files. The algorithms try to look for reads that map exactly in the same reference location.

SNP calling

One of the main applications of the NGS technologies is the SNP mining in the resequencing projects. Reads from different individuals are generated and Single Nucleotide Polymorphisms (SNPs) and indels are looked for by comparing them with the reference genome.

Once an alignment is generated as a BAM file looking for SNPs is not a conceptually a difficult task. We go through every column of the alignment and in every one we see how many alleles are found and how they compare with the one found in the reference genome. Unfortunately this naive view is complicated by several confounding factors:

- The cloning process artifacts (e.g. PCR induced mutations).
- The error rate associated with the sequence reads.
- The error rate associated with the mapping.
- The reliability of the reference genome.

In fact Heng Li, the author of BWA, has recently evaluated the error rate of the SNP calling proccess and has concluded that the two major sources of errors are:

- erroneous realignment in low-complexity regions
- incomplete reference genome with respect to the sample

He concludes that with the methods available at April of 2014 "the raw genotype calls is as high as 1 in 10-15 kb, but the error rate of post-filtered calls is reduced in 1 in 100-200kb without significant compromise on the sensitivity".

Alignment considerations

The mapping tools calculate a probability for the correctness of the alignment for the whole read. This probability depends on the length of alignment, on the number of mismatches and gaps and on the uniqueness of the aligned region on the genome and it should reflect the probability of the read being originate from the aligned region on the reference. It is important to distinguish the real SNPs from the mismatches between repeated homologous genomic regions.

Even in the case in which the read maps only to one location in the reference genome and we have a good alignment score for the overall read some bases of the read can be misaligned.

```
coor     12345678901234    567890123456789012345678
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaac****aaAtaa
```

```
read2    ggttttataaaac****aaAtaaTt
read3       ttataaaacAAATaattaagtctaca
read4          CaaaT****aattaagtctacagagcaac
read5            aaT****aattaagtctacagagcaact
read6              T****aattaagtctacagagcaacta
```

One approach to this problem is to realign the problematic regions to solve the problem, this is the approach taken by GATK realignment. The actual implementations of this realignment are computationally quite intensive and the results are not perfect. The samtools developers have proposed an alternative solution, instead of solving the problem, to detect it and mark it with alignment qualities per base and not only per read. The resulting qualities calculated by the samtools are known as BAQ (Base Alignment Quality) and the method to calculate them is described in the mpileup manual.

Quality recalibration

Every base of the reads is generated with a Phred score associated. This score should be related with the probability of a sequencing error on the nucleotide read. In this way we could distinguish sequencing errors from real variation, but there is a catch, the Phred values have an intrinsic error in themselves. When the Phread values are compared with the real sequencing error rates, calculated by resequencing well established standards, they are usually found to be in disagreement. It is often the case that the sequencing error rates predicted by the sequencing machines are not completely accurate. To solve this issue a recalibration of the quality scores can be carried out.

To do a recalibration the variable positions found in the alignments are classified according to the information of which SNPs have been previously detected in the species. The variable positions that do not match a previously known SNP are expected to be mainly sequencing errors and with that information the read quality can be recalibrated. This process is implemented by GATK and SOAPsnp.

In the case of a species with much previous SNP information the recalibration could be carried out by doing a first round of SNP calling and then recalibrating using the called SNPs as the true SNP of the species. In this case after the recalibration is done the second, and definitive, SNP calling would be performed.

SNP calling

Once we have taken into account the sequencing and alignment problems we can use a SNP calling software to look for the SNPs. The most commonly used SNP callers are: samtools' mpileup, GATK and FreeBayes. Each one of these SNP callers make different assumptions about the reference genome and the reads, so each one of them is best suited for different situations.

Some SNP callers are based on counting the number of reads for each alleles once appropriate thresholds for the sequencing and mapping qualities have been applied. This

simple method is the one used by the VarScan SNP caller as well as by most of the commercial SNP callers. But other methods based on more advanced statistics have also been developed. This methods often perform better, specially with low coverages, and do certain assumptions to create bayesian models. Most of them assume diploid individuals and some even take into account the Hardy-Weinberg equilibrium and Linkage Disequilibrium information as well as previous information about the SNPs present in the species and their allele frequencies.

The GATK project has published a good resource to lear more about SNP calling best practices.

Brad Chapman has a very interesting piece comparing the use of several aligners and SNP callers. For the read alignment he used bwa-mem. Then he compared two alternative post-processing methods:

- de-duplication with Picard MarkDuplicates, GATK base quality score recalibration and GATK realignment around indels.
- Minimal post-processing, with de-duplication using samtools rmdup and no realignment or recalibration.

For the SNP and indel calling with compared three methods:

- FreeBayes (v0.9.9.2-18): A haplotype-based Bayesian caller from the Marth Lab.
- GATK UnifiedGenotyper (2.7-2): GATK's widely used Bayesian caller.
- GATK HaplotypeCaller (2.7-2): GATK's more recently developed haplotype caller which provides local assembly around variant regions

Some of his main conclusions were:

- skipping base recalibration and indel realignment had almost no impact on the quality of resulting variant calls
- FreeBayes outperforms the GATK callers on both SNP and indel calling. The most recent versions of FreeBayes have improved sensitivity and specificity which puts them on par with GATK HaplotypeCaller.
- GATK HaplotypeCaller is all around better than the UnifiedGenotyper.

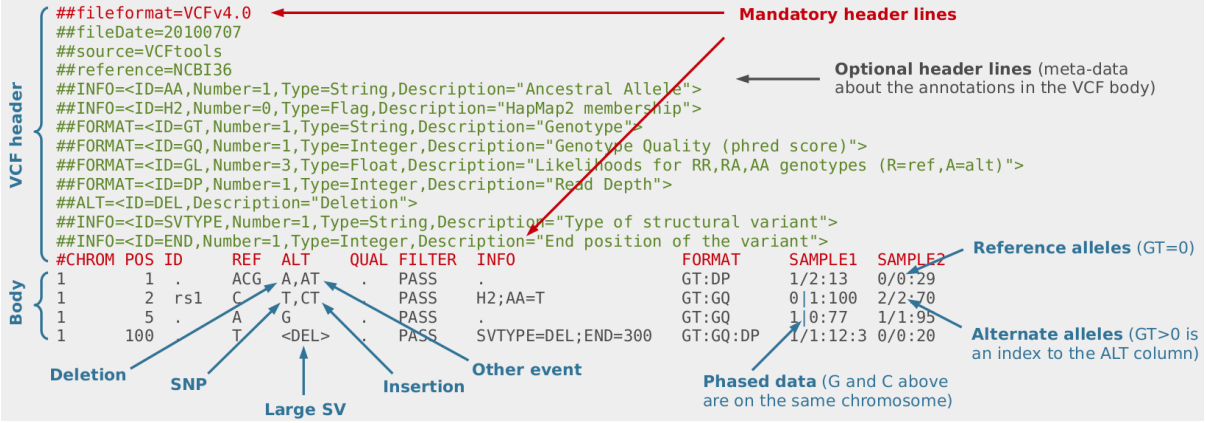
He has also compared the performance of the Structural variant callers and the cancer SNP callers.

VCF format

The end result of a SNP calling analysis is a collection of SNPs. An standard file has been created to hold these SNPs, the Variant Call Format file (VCF). In this file every line represents an SNP and the following information is found:

- The position in the reference genome.
- The allele in the reference genome.

- The other alleles found.
- The filters not passed by the SNP.
- The genotypes found with its abundances.



A tool for working with these files has been created, VCFtools. In its web site a definition of the file format can be found. VCFtools allow:

- Format validation.
- SNV annotation.
- VCF comparison.
- Statistics calculation.
- Merging, intersections and complements.

We can take a look at a VCF file quite easily with a text editor, although we might also find some of them convected to a binary format (BCF). Also, the fields in this file are delimited by tabs, so it can be imported into a spreadsheet program by using the csv option.

The types of variants that can be stored in a VCF file are:

SNPs
Alignment VCF representation
ACGT POS REF ALT
AtGT 2 C T

Insertions
Alignment VCF representation
AC-GT POS REF ALT
ACtGT 2 C CT

Deletions
Alignment VCF representation
ACGT POS REF ALT
A--T 1 ACG A

Complex events
Alignment VCF representation
ACGT POS REF ALT

A-tT 1 ACG AT

Large structural variants
VCF representation
POS REF ALT INFO
100 T <DEL> SVTYPE=DEL;END=300

SNP quality assessment

Several distributions can be created to analyze the SNP calling result:

- % of missing calls per SNP
- SNP depth or Genotype depth
- SNP observed heterozygosity
- SNP quality
- SNP density
- Sample depth
- % of missing calls per sample
- Sample observed heterozygosity

SNP filtering

Filtering the SNPs after the SNP calling is a critical task. We can filter the SNPs for different reasons like usefulness or risk of being a false positive. In the called SNPs there will be some false positives so we could want to remove those false positives. It is common to divide the SNPs in several tiers according to our confidence in them.

Several application exist to filter SNPs VCFtools, SnpSift, Vardict and GATK are just some examples.

Some of the parameters than can be taken into account are: quality, heterozygosity, depth, mapping quality, errors of the reads, or allele frequency.

We could also select some SNPs for a genotyping platform or to do a particular analysis.

A VCF file is a matrix with the SNPs in rows, the samples (e.g. individuals) in the columns and the genotypes in the cells. We can filter SNPs (lines/rows), samples (columns) or genotypes (setting the corresponding genotype to not determined). In the case of the SNP filtering the nomenclature can be confusing, because two different kind of analyses are commonly refered as filtering. We can remove the lines corresponding to the filtered SNPs from the file altogether or we can annotate the SNP/row adding a tag to the filter column in the VCF file, but without removing the SNP from the file.

The freebayes SNP caller includes some programs to filter the SNPs, among them vcffilter that makes possible to remove SNPs/rows or genotypes from the VCF file by using different criteria.

Filter for SNPs

Low quality

SNP callers usually assign a quality (probability) to the SNPs. We can filter out the SNPs with lower qualities.

Missing data

We could filter the SNPs with large amount of missing genotypes. This could happen, for example, in RNASeq experiments (in genes with low expression in some samples), in GBS experiments or in low coverge genome sequencings.

Number of alleles

It is possible to remove the monomorphic SNPs or to filter out the SNPs that are not biallelic.

Kind

We can filter the SNVs according to its type: SNV, indel, complex or structural variation

Position

We can filters the SNPs according to its location in the genome. For instance, we could keep only the SNPs found in an exon or in a coding region.

It is also common to thin out the SNPs, to select one SNP every some kilobases in the genome.

Low Complexity Region

It has been shown that due to problems with the PCR and the alignment the low complexity regions are particularly prone to false positive SNPs. We could remove them with a low complexity filter. These are also the regions that tend to be more variable in the populations, so by removing those SNPs we will create lots of false negatives. This filter will tend to decrease the amount of information, but hopefully will also remove quite a lot of noise.

Flag and info

We could filter the SNPs according to the flag and info fields found in the VCF files. It is usual that a tools that runs a filter in a VCF file just puts a tag in the VCF flag field.

Minor Allele Frequency (MAF)

MAF can sometimes refer to the Minor Allele Frequency and sometimes to the Major Allele Frequency. Both statistics convey the same information for the biallelic SNPs, but the Major Allele Frequency is more straightforward if we have more than 2 alleles.

SNPs due to sequencing errors will usually have major allele frequencies close to 1, because few genotypes will have an allele due to the error. So we could remove most SNPs due to sequencing errors by using this filter. If we do it, we will also filter out lots of real SNPs that are almost fixed in the population.

If we are dealing with a segregant population we usually expect a range of MAF values and we can use this information to decide which SNPs should be filtered out.

If we have pooled samples we might consider applying this filter to individual samples.

Another very related measure is MAC: major/minor allele count.

Observed Heterozygosity

One common source of false positive SNPs with high heterozygosity rates is due to duplicated regions found in the problem sample that are not found in the reference genome. It is common to have SNPs in these regions with heterozygosities close to 0.5. In such cases the SNPs will be due to reads from the two copies that are piled in the only copy found in the reference genome. This cases can not be avoided by filtering the reads with MAPQ because since only one copy of the duplication is found in the reference genome the mapper software can not guess that there is a problem due to a repetitive element. Another way to spot these false positives is to look for SNPs with a high coverage.

High Coverage

An excessive coverage can point to false positives due to duplicated regions in the sequenced sample not found in the reference genome. See also the observed heterozygosity filter.

Highly Variable Region

Having regions with too many SNPs could also be a sign that we are piling up reads from repeated regions. We could filter out the SNPs located in such highly variable regions. This analysis is usually done counting the number of SNPs in a window around each SNP.

It can also be useful to remove the SNPs with an SNP too close if we want to design primers to do a PCR or genotyping experiment. In this case we might also want to remove the SNPs that are close to the start or the end of the reference sequence. This could be particularly relevant if we are using a transcriptome as a reference.

Linkage Disequilibrium

If we have genotype a segregant population it could be useful to filter out the SNPs that are not in linkage disequilibrium with their closest SNPs. Many of these unlinked SNPs will be false positives.

Variability

We might be interested in filtering out or selecting SNPs that are variable in a set of samples or that differenciate two sets of samples.

Aminoacid change

We can select the SNPs with large impacts in the coded proteins. The SnpEff tool can be used for that.

Cap enzyme

We can select the SNPs that create restriction sites if we want to detect them by PCR and restriction enzyme digestion.

HWE

We can also filter out the SNPs that are not in HWE or that show a non-medelian segregation in a segregant population.

Filters for Genotypes

It is also possible to filter out not SNPs, but genotypes. In this case the genotype is usually set to not determined.

To genotype a sample with good quality we need more information than to just get the SNP with good quality. If we have several samples, all their reads will contribute information to determine the SNP, but to get the genotype of any of them we need enough coverage in the given sample.

Two common filters used for genotypes are the depth of coverage for the genotypes and the genotype quality that is created by most SNP callers.

GFF format

The GFF files are used to store annotations. An annotation can be thought as a label applied to a region of a molecule. For instance we could tag a region covered by a gene in a chromosome. The GFF files are text files and every line represents a region on the annotated sequence and these regions are called features. In the previous case the gene would be a feature of the chromosome. Features can be functional elements (e.g., genes), genetic polymorphisms (e.g. SNPs, INDELs, or structural variants), or any other annotations. Each feature should have a type associated. Examples of some possible types are: SNPs, introns, ORFs, UTRs, etc. The terms used to define these types should belong to the Sequence Ontology terms. If you are interested you can take a look at the Sequence Ontology or at the GFF format specification.

In the GFF format both the start and the end of the features are 1-based.

```
##gff-version 3
##sequence-region ctg123 1 1497228
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
```

BED format

The BED format provides a simpler way of representing the features in a molecule. Each line represents a feature in a molecule and it has only three required fields: name, start and end.

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000
```

The BED format uses 0-based coordinates for the starts and 1-based for the ends. So the 1st base on chromosome 1 would be:

```
chr1    0    1    first_base
```

Headers are allowed. Those lines should be preceded by # and they will be ignored.

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 – II - NEXT GENERATION SEQUENCING – SBI1606

Whole genome sequencing

The aim of whole-genome sequencing (WGS) is to determine an organism's complete DNA sequence in a single experiment, including a comprehensive picture of both the coding and non-coding regions. As such, WGS provides a comprehensive picture of both the coding and noncoding regions of chromosomal and mitochondrial DNA, as well as chloroplast DNA (in plants). WGS enables the detection of all types of genetic variation, including single-nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and structural variants, such as translocations and copy number variation (CNV).[1]

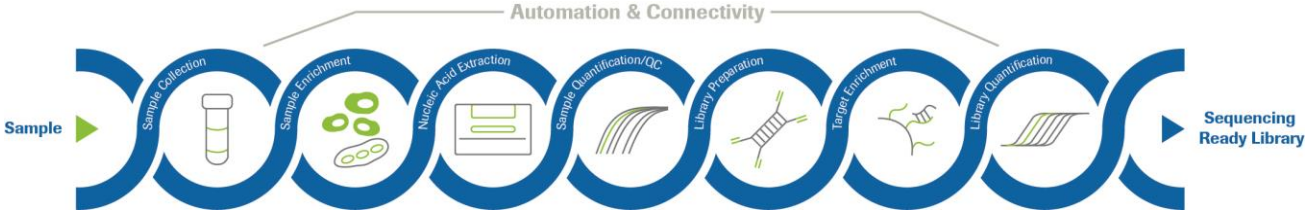
The genome of bacteriophage ϕX174 (5,386 bp) was the first genome to be fully sequenced, by Fred Sanger and colleagues in 1977.[2] In the 14 years that followed, the Sanger method was used to sequence small genomes, such as those of bacteriophages and viruses (all in the 50 – 200 kb range); as well as the first genome of a free-living organism (Haemophilus influenza, 1.8 Mb; published in 1995[3]). Sanger sequencing was also used to sequence the first plant genome (Arabidopis thaliana, 135 Mb; published in 2000[4]) and the first draft of the human genome (published in 2001[5]). The advent of next-generation sequencing (NGS) made sequencing of the first human cancer genome possible (published in 2008[6]). Continuous improvements in NGS technology (and concomitant reductions in per-base cost) have since enabled routine, high-throughput WGS of both simple and highly complex genomes.

Among other applications, WGS research enables us to:

- gain deeper insight into the genomic basis of health, disease and ancestry than what is possible with targeted sequencing approaches
- discover biomarkers and understand pharmacogenetics
- perform genome-level comparative analysis, to identify synteny, orthologs and horizontal gene transfer events
- generate reference genomes for agriculturally important animals and plant, to assist with breeding
- support ecology and conservation biology
- understand disease outbreaks and public health
- secure food safety
- understand antibiotic resistance
- study microbiomes and their role in human health and disease

Sample Prep for whole-genome sequencing

As is the case for all NGS applications, sample prep constitutes the first step in the WGS workflow, and holds the key to unlocking the potential of every sample. Because NGS samples are precious, the best sample prep solutions are needed to process more samples successfully, get more information from every sample and optimize your sequencing resources. Roche Sample Prep Solutions offer an integrated approach to sample preparation, addressing all of the steps required to convert a sample to a sequencing-ready library. From sample collection to library quantification, we offer sample prep solutions for different sample types and sequencing applications that are proven, simple and complete.



Library construction for WGS starts with fragmenting DNA to the appropriate size, after which platform-specific adapters are added. PCR-free workflows are preferred for WGS, but in cases where input DNA is limited or is of poor quality, library amplification is required. WGS library construction protocols typically include a size-selection step as a narrow library fragment distribution facilitates data analysis. Quantification and QC of sequencing-ready libraries are important to ensure optimal clonal amplification on NGS platforms. After sequencing, sequence reads are aligned against a reference genome (reference-guided sequence assembly), or when no such reference is available, compared to each other and assembled into long contiguous segments (de novo sequencing). This general workflow applies to the sequencing of both simple (e.g. bacterial) and complex (e.g. human) genomes, but these applications pose very different challenges.

Advantages of Whole-Genome Sequencing

- Provides a high-resolution, base-by-base view of the genome
- Captures both large and small variants that might be missed with targeted approaches
- Identifies potential causative variants for further follow-up studies of gene expression and regulation mechanisms
- Delivers large volumes of data in a short amount of time to support assembly of novel genomes

An Uncompromised View of the Genome

Unlike focused approaches such as exome sequencing or targeted resequencing, which analyze a limited portion of the genome, whole-genome sequencing delivers a comprehensive view of

the entire genome. It is ideal for discovery applications, such as identifying causative variants and novel genome assembly.

Whole-genome sequencing can detect single nucleotide variants, insertions/deletions, copy number changes, and large structural variants. Due to recent technological innovations, the latest genome sequencers can perform whole-genome sequencing more efficiently than ever.

Introduction to Large Whole-Genome Sequencing

Sequencing large genomes (> 5 Mb) can provide valuable information for disease and population-level studies. Researchers often use large whole-genome sequencing to analyze tumors, investigate causes of disease, select plants and animals for agricultural breeding programs, and identify common genetic variations among populations.

Advantages of Large Whole-Genome Sequencing

- Provides a high-resolution, base-by-base view of the genome
- Combines short inserts and longer reads to allow characterization of any genome
- Reveals disease-causing alleles that might not have been identified otherwise
- Identifies potential causative variants for further follow-on studies of gene expression and regulation mechanisms

A Comprehensive View of Genetic Variation

Analyzing the whole genome using next-generation sequencing (NGS) delivers a base-by-base view of all genomic alterations, including single nucleotide variants (SNV), insertions and deletions, copy number changes, and structural variations. Paired-end whole-genome sequencing involves sequencing both ends of a DNA fragment, which increases the likelihood of alignment to the reference and facilitates detection of genomic rearrangements, repetitive sequences, and gene fusions.

Introduction to Small Whole-Genome Sequencing

Small genome sequencing (≤ 5 Mb) involves sequencing the entire genome of a bacterium, virus, or other microbe, and then comparing the sequence to a known reference. Sequencing small microbial genomes can be useful for food testing in public health, infectious disease surveillance, molecular epidemiology studies, and environmental metagenomics.

Advantages of Small Genome Sequencing

- Allows investigation of all genes from single organism culture
- Sequences thousands of organisms in parallel
- Provides comprehensive analysis of the microbial or viral genome
- Aids discovery of new biomarkers within a microbial or viral sample by providing distinct gene information from homologous chromosomes, supporting haplotyping

Fast, Culture-Free Microbial Analysis

Unlike traditional approaches, small whole-genome sequencing (WGS) studies using next-generation sequencing (NGS) do not rely on labor-intensive cloning steps. NGS also enables biologists to sequence hundreds of organisms simultaneously via multiplexing. NGS can identify low-frequency variants, genomic rearrangements, and other genetic changes that might be missed or are too costly to identify using other methods. For small genomes, DNA libraries can be prepared, sequenced, and analyzed in as little as 2 days.

What Is* De Novo *Sequencing?

De novo sequencing refers to sequencing a novel genome where there is no reference sequence available for alignment. Sequence reads are assembled as contigs, and the coverage quality of *de novo* sequence data depends on the size and continuity of the contigs (ie, the number of gaps in the data).

Next-generation sequencing (NGS) allows faster, more accurate characterization of any species compared to traditional methods, such as Sanger sequencing. Illumina NGS technology offers rapid, comprehensive, accurate characterization of any species.

Advantages of* De Novo *Sequencing

- Generates accurate reference sequences, even for complex or polyploid genomes
- Provides useful information for mapping genomes of novel organisms or finishing genomes of known organisms
- Clarifies highly similar or repetitive regions for accurate *de novo* assembly
- Identifies structural variants and complex rearrangements, such as deletions, inversions, or translocations

Accurate* De Novo *Genome Assembly

When sequencing a genome for the first time, a combined approach can yield higher-quality assemblies. For example, combining short-insert, paired-end and long-insert, mate pair sequences is an ideal way to maximize coverage. The short reads, sequenced at higher depths, can fill in gaps not covered by the long inserts.

This combination enables detection of a broad range of structural variant types and is essential for accurate identification of complex rearrangements.

Introduction to Phased Sequencing

Historically, whole-genome sequencing generated a single consensus sequence without distinguishing between variants on homologous chromosomes. Phased sequencing, or genome

phasing, addresses this limitation by identifying alleles on maternal and paternal chromosomes. This information is often important for understanding gene expression patterns for genetic disease research.

Benefits of Phased Sequencing

Next-generation sequencing (NGS) enables whole-genome phasing without relying on trio analysis or statistical inference. By identifying haplotype information, phased sequencing can inform studies of complex traits, which are often influenced by interactions among multiple genes and alleles. Phasing can also provide valuable information for genetic disease research, as disruptions to alleles in cis or trans positions on a chromosome can cause some genetic disorders.

Phasing can help researchers to:

- Analyze compound heterozygotes
- Measure allele-specific expression
- Identify variant linkage

Applications of WGS: Case Studies

De novo Genome Assembly

Generally a genome is assembled from NGS sequence data by aligning to a reference genome. For most organisms however, this is not possible since no reference is available. In these cases *de novo* genome assembly is preformed.

When assembling a genome without a reference it is essential to have a way to correlate sequences long range, otherwise the assembly will be inaccurate. There are two ways to achieve this. Often a library of fosmids is constructed and sequenced by Sanger Sequencing in parallel to NGS. While sequencing the fosmids does not provide the coverage of NGS, it gives very long reads and so allows the correlation of sequences that are far from each other. The drawback is that the large amount of Sanger Sequencing required is both expensive and time consuming. More recently a new approach has been taken that relies entirely on illumina's short sequencing reads. Here many libraries are generated, some with very long lengths. The libraries are then undergo paired-end sequencing to generate mate-paired sequences with correlated short reads and a gap of known length but unknown sequence. These mate-pairs allow long range correlation between sequences, allowing a more accurate genome assembly.

This approach was used recently to sequence the genome of the flax plant (*Linum usitatissimum*) by an international collaboration. Flax is an important crop for both food and textile production. Sequencing its genome will help agronomists develop better varieties and

better understand the domestication of this crop. The authors generated seven libraries with varying lengths from 300 bp to 10 kb and sequenced them using paired-end illumine (**Figure 1**). This generated mate-pair and paired-end reads with 44-100 bp of known sequence and a spacer of defined length (**Figure 1**). The use of mate-paired reads with thousands of bases between them allowed the alignment of sequences long range, enhancing the accuracy of the assembly.

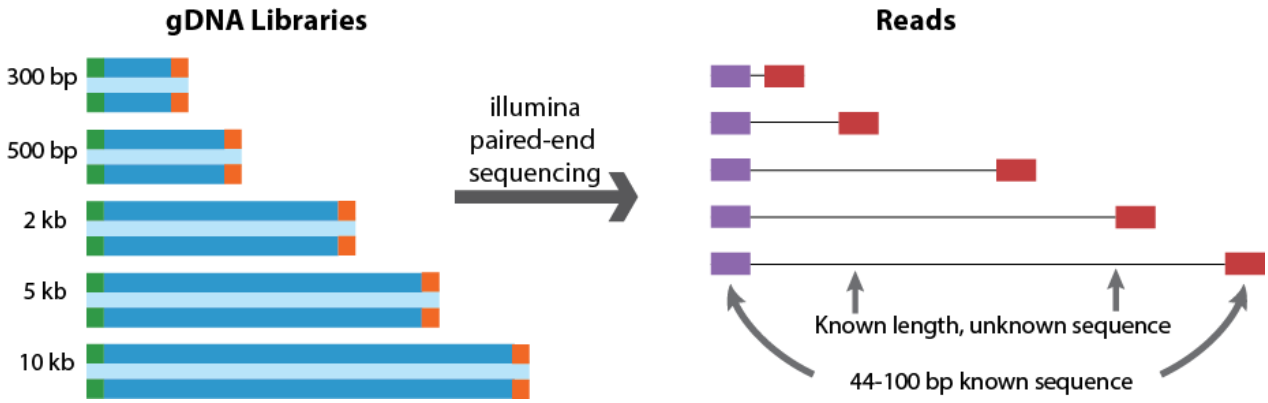


Figure 1 – For flax genome assembly libraries of 300 bp to 10 kb were prepared. These were sequenced as paired-end reads

The first step in assembly was to remove low quality reads, after which the coverage was determined to be 69x. After filtering, the reads were aligned to each other to generate 116,602 contigs (**Figure 2**). The contigs were further aligned to generate 88,384 scaffolds, 132 of which contained 50% of the assembly and were longer than 693.5 kb (**Figure 2**). The longest scaffold was 3.09 Mb. The assembly was found represent 85% of the genome with an average of 45x coverage.

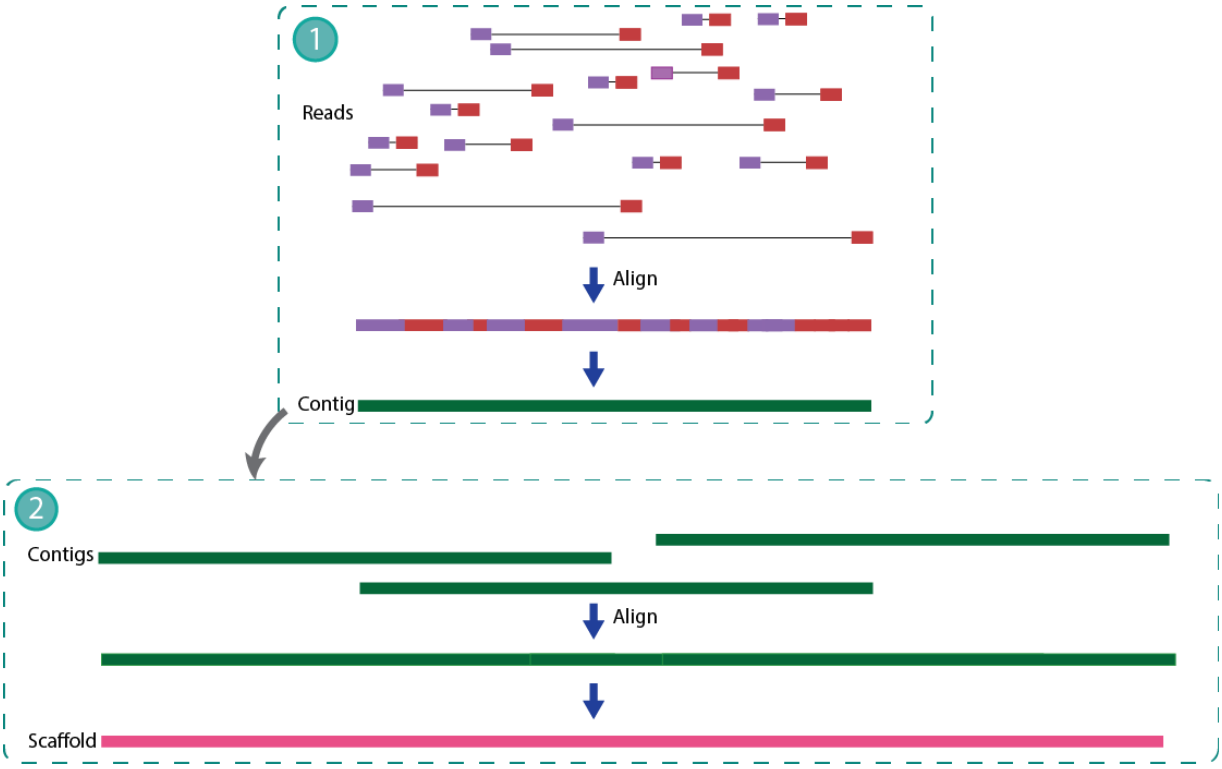


Figure 2 – Reads are aligned to make contigs and contigs are then aligned to make scafolds.

Expressed sequence tags (ESTs) are short sequences obtained from cDNA libraries. They represent expressed regions of the genome and so can be used to find genes. In this study the known ESTs of flax were aligned to the scaffolds. Ninety-three percent of the flax ESTs aligned to the WGS scaffolds with >95% sequence identity indicating the assembled genes were highly accurate. This study also preformed many different analyses including comparison of the assembled flax genome to the genome of other plants. For more information please see the original paper.

Pathogen Tracking

Whole Genome Sequencing can also be used to track pathogen outbreaks. At present, the gold standard for analyzing strains of pathogenic bacteria is pulsed-field gel electrophoresis (PFGE), which compares the banding pattern between genomes digested by a selected restriction enzyme. This approach is limited however, since significant mutations can be easily hidden when they don't affect the restriction sites or relative size of the genomic

fragments. At the same time a single nucleotide mutation can result in the gain or loss of a restriction site and so can give a different PFGE pattern between closely related strains. This proof of concept study by Revez et al investigates WGS as a replacement for PFGE.

Campylobacter jejuni is among the leading causes of food born illness in the world. It is naturally found in the guts of birds and cows. Humans are most likely to become infected by injecting contaminated water. *C. jejuni* infection is debilitating, but rarely fatal. In this study samples from a *Campylobacter jejuni* outbreak in Europe in 2012 were reanalyzed by WGS and compared to the conclusions drawn from standard methodologies, to decide if WGS has similar or enhanced ability to track pathogen source and evolution during an outbreak situation.

Based on the PGFE patterns observed during the outbreak it was concluded that there was a contamination event involving one strain and one water source. However WGS revealed that this was not the case (**Figure 3**). Of the two human isolates shown here, one was found to be highly similar to the waterborne strain. The other human isolate is highly divergent, too much so to be the result of genetic drift during the course of the outbreak. In light of the WGS data the authors conclude that either a single source of water was contaminated by multiple divergent strains or that there were multiple sources of contamination. These results highlight the importance of more accurate WGS data during pathogen outbreaks, since conventional methodology misidentified a patient strain, potentially missing other sources of contamination.

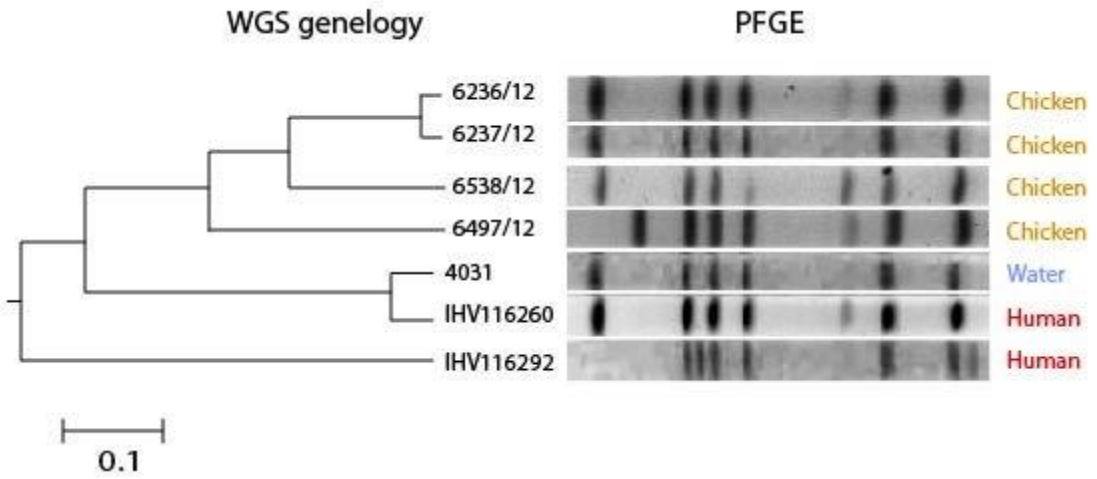


Figure 3 – The relationships between strains determined by WGS was more accurate than those that could be observed by PFGE. For example IHV116260 and 6237/12 are indistinguishable by PFGE but WGS revealed that they are highly divergent.

Molecular Evolution

Whole genome sequencing is also an essential tool for studying molecular evolution. This study uses WGS to study the molecular evolution of the Ithica New York honeybee population in response to the introduction of the mite *Varroa destructor*. Specimens collected in 2010 were compared to museum specimens collected in 1977, before the introduction of the mite. Honeybees, *Apis mellifera*, are essential to human agriculture. Both feral and domestic populations exist in North America. Honey bees are a eusocial species; each colony contains a sexually mature queen bee, a few thousand haploid males and tens of thousands of sterile female worker bees (**Figure 4**). The mite *Varroa destructor* feeds on the hemolymph of the adult worker bees, weakening them and making them more susceptible to disease (**Figure 4**). It has been associated with colony collapse.

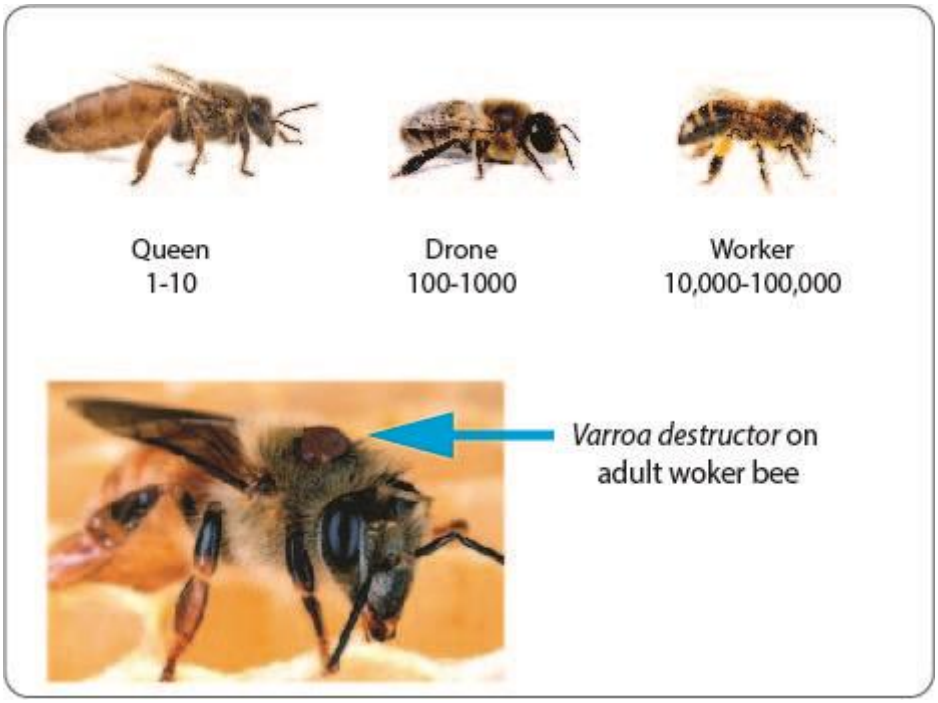


Figure 4

The authors found a drastic loss of mitochondrial haplotypes between 1977 and 2010, with an entire clade going extinct (**Figure 5**). This loss indicates a population bottleneck upon the introduction of *Varroa destructor*. However they also found no decrease in nuclear

genetic diversity. This finding indicates that the modern population is descended from a small number of queens through high rates of outbreeding and polyandry. The ancestry of the modern bee population is similar to the museum bees with a few variants. In the modern bees there is traces of African and Arabian ancestry that was absent in the museum bees (**Figure 5**). The authors also found some genes that were under selection pressure in the modern population relative to the museum bees which may play a role in resistance to the *Varroa destructor* parasite. For more details please see the original paper. This study demonstrates that WGS is a powerful tool for studying the molecular evolution of a population over time.

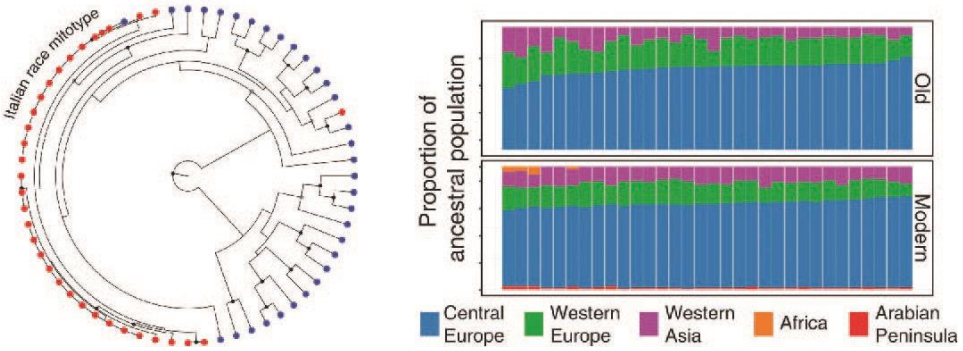


Figure 5

Whole-Genome Sequencing Methods

Sequencing technologies are unable to sequence the entire human genome at once. Thus, the genome must be broken into smaller chunks of DNA, sequenced and then put back together in the correct order using bioinformatics approaches. There are several methods of DNA sequencing, including clone-by-clone and whole-genome shotgun methods. For more information on whole-genome sequencing as it relates to field of immuno-oncology, see the section "Types of Molecular Testing -- Research."

Clone-by-clone

This method requires the genome to have smaller sections copied and inserted into bacteria. The bacteria then can be grown to produce identical copies, or “clones,” containing approximately 150,000 base pairs of the genome that is desired to be sequenced. Then, the inserted DNA in each clone is further broken down into smaller, overlapping 500 base pair chunks. These smaller inserts are sequenced. After sequencing is performed, the overlapping portions are used to reassemble the clone. This approach was used to sequence the first

human genome using Sanger sequencing. This approach is time-consuming and costly, but it is reliable.

Whole-genome shotgun

As the name implies, "shotgun" sequencing is a method that breaks DNA into small random pieces for sequencing and reassembly. The pieces of DNA are also cloned into bacteria for growth, isolation and subsequent sequencing. Because the pieces are random, there are overlapping sequences that aid in reassembly into the original DNA order. This approach was originally used in Sanger sequencing but is now also used in next-generation sequencing methods providing rapid genome sequencing with lower costs. It is only good for shorter "reads" (ie, sequencing on shorter DNA fragments to be put back together again). Because it is reassembled based on overlapping regions and has shorter read lengths, it is best utilized when a reference genome is available, and it requires sophisticated computational approaches to reassemble the sequence. It also can be challenging for genomes with many repetitive regions.

Assembly of sequencing reads

Because genomes are sequenced in varying lengths of DNA fragments, the resulting sequences must be put back together. This is referred to as "assembly," or "reassembly." Two common approaches are de novo assembly and assembly by reference mapping.

De novo assembly is performed by identifying overlapping regions in the DNA sequences, aligning the sequences and putting them back together to form the genome. This is done without any sequence with which to compare. Mapping to a reference genome uses another genome to align new sequencing data to as a comparator.

Although de novo assembly can be challenging, this approach is the only one available for sequencing new organisms. Additionally, de novo assembly introduces results with less bias than mapping to a reference genome. Mapping to a reference genome is easier and requires less contiguous reads, but new or unexpected sequences can be lost. The sequence results obtained by this method is only as good as the reference genome chosen; however, it can provide better identification of single nucleotide polymorphisms (SNPs). Multiple institutions and genomic sequencing companies have invested considerable time and effort into creating improved reference genomes. Single nucleotide polymorphisms are known to vary by race and ethnicity, thus, multiple reference genomes have been created for various races/ethnicities.

Examples of next-generation sequencing platforms

Several companies focus on development and marketing of next-generation sequencing machines (often referred to as "platforms") for use in whole-genome (and other) sequencing. Illumina is considered by many as the leader because of the number of users that utilize its systems. Illumina has multiple platforms depending on the need. The Illumina HiSeq is one of the more common sequencers found in laboratories, including major research institutions, companies providing next-generation sequencing services for clinics and labs, and pathology laboratories. It has a high throughput, capable of sequencing many genomes rapidly with reasonable costs. This instrument also can be used to look at copy number variation, as well as mutations and other alterations, and RNA expression levels to do transcriptomics. Because of the popularity in the clinic of targeted sequencing panels, which are much smaller with clinics requiring faster turnaround times for treatment of patients, Illumina created the MiSeq, which can provide same-day sequencing results for very small panels. Illumina also produced multiple variations to provide sequencers for each disease area optimizing output, turnaround time and costs for specific use cases.

Thermo Fisher Scientific's Ion Torrent or Ion Proton uses a completely different technology based on detection of pH differences and was once expected to provide better utility for clinical applications because it was easier to use, cost less and provided faster turnaround time. However, Illumina countered with new machines to fit these needs. Consequently, both are found in research and clinical laboratories.

Other technologies developed recently use different novel approaches. A few examples are provided below.

Oxford Nanopore Technologies introduced the MinION, which enables anyone to sequence on a desktop computer using a USB device. The DNA is passed through a protein nanopore membrane for sequencing and detection by creation of an ionic current that varies based on the nucleotide.

Pacific Biosciences introduced its single molecule, real-time technology with the longest reads to date, with average read lengths of more than 10,000 base pairs compared with more than 150 base pairs. Single molecule, real-time technology uses a chip with single DNA molecules attached. Zero-mode waveguide technology enables isolation of a single nucleotide for the DNA polymerase to add fluorescent labels for detection of each base. The error rate of this instrument is still higher than some of the prior technologies, but a lot of

interest has been generated, and there is hope that speed and costs can be further optimized with the new approach.

Coverage breadth and depth

Coverage refers to the number of reads that show a specific nucleotide in the reconstructed DNA sequence. A read is a string of A, T, C, G bases that correspond to the reference DNA. There are millions of reads in a sequencing run. Increased coverage depth results in increased confidence in variant identification.

For the human genome, a 10- to 30-times coverage depth is acceptable for detecting mutations, SNPs and rearrangements. A next-generation sequencing approach that provides a coverage depth of 30 times is considered to have high coverage. However, as coverage depth increases, coverage breadth decreases (Figure 1).

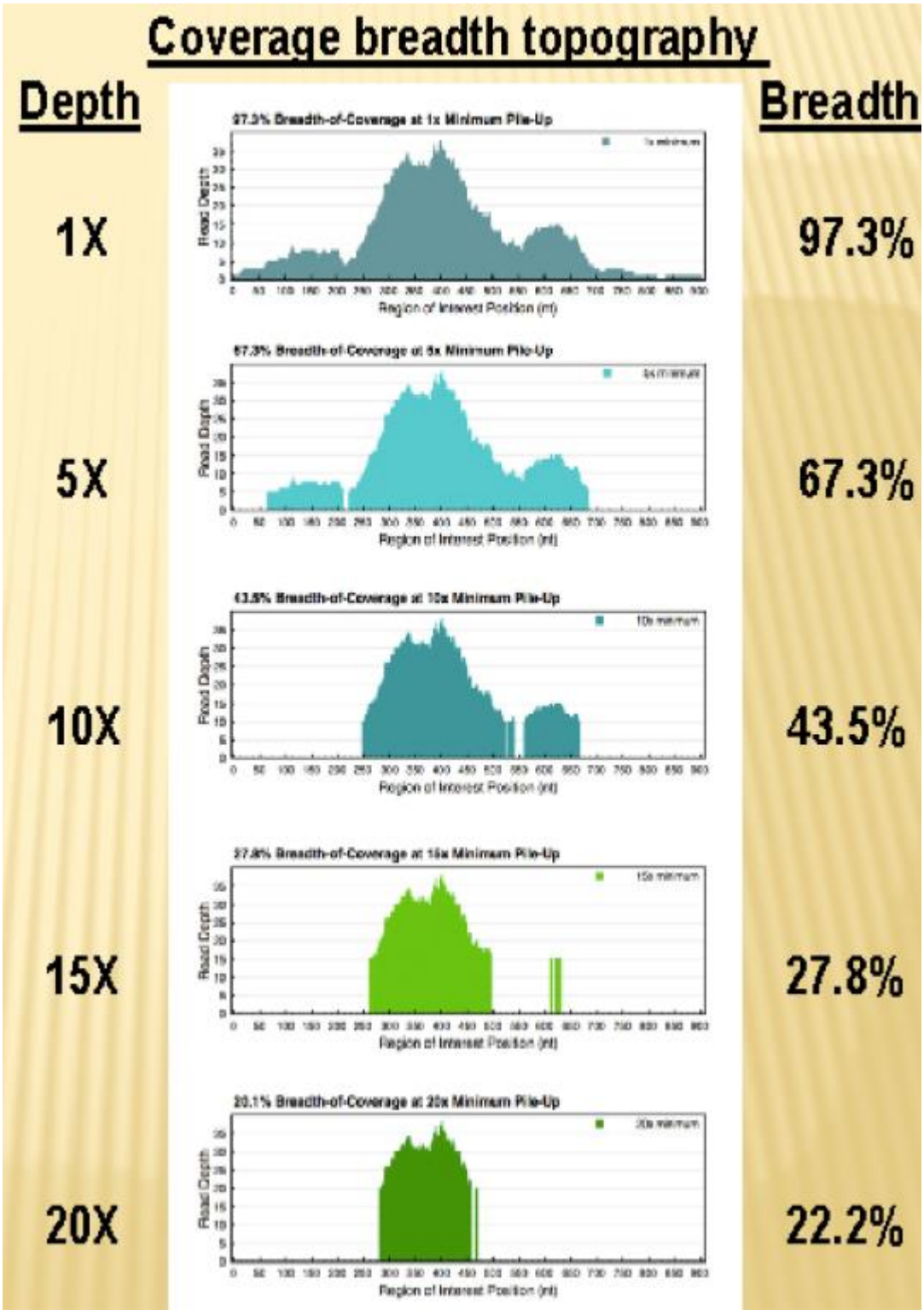


Figure 1. Relationship between coverage breadth vs. coverage depth.

Whole-Genome vs. Whole-Exome Sequencing vs. Targeted Sequencing Panels

Whole-genome sequencing determines the order of the nucleotides (A, C, G, T) in the entire genome that makes up an organism. The goal of whole-genome sequencing is, typically, to look for genetic aberrations (eg, single nucleotide variants, deletions, insertions and copy number variants). Because the entire genome is being sequenced, changes in the noncoding sections of DNA within genes, called introns, can also be determined. Under normal conditions, introns are removed by RNA splicing during a post-transcriptional process, and alterations in these regions can be important to whether the DNA is transcribed into RNA or potentially result in a truncated, non-functional protein.

An alternative approach is to sequence only the exomes, called whole-exome sequencing. Exomes are the part of the genome formed by exons, or coding regions, which when transcribed and translated become expressed into proteins. Exomes compose only about 2% of the whole genome. Because the genome is so much larger, exomes are able to be sequenced at a much greater depth (number of times a given nucleotide is sequenced) for lower cost. This greater depth provides more confidence in low frequency alterations. Sequencing depth can become even greater for lower cost by using a targeted or “hot-spot” sequencing panel, which has a select number of specific genes, or coding regions within genes that are known to harbor mutations that contribute to pathogenesis of disease, and may include clinically-actionable genes of interest (eg, diagnostic, theranostic, etc.). These are often used in clinical care to provide greater confidence as well as keep the cost down and provide better opportunity for insurance reimbursement. However, whole-exome sequencing and targeted panels only see part of the story as they focus on reduced areas of the genome. Consequently, for some research projects or genetics testing, whole-genome sequencing may be advantageous.

Strengths and Limitations of Next-Generation Sequencing

Strengths

The major strength of next-generation sequencing is that the method can detect abnormalities across the entire genome (whole-genome sequencing only), including substitutions, deletions, insertions, duplications, copy number changes (gene and exon) and chromosome inversions/translocations. A major strength of next-generation sequencing is that it can detect all of those abnormalities using less DNA than required for traditional DNA sequencing approaches. Next-generation sequencing is also less costly and has a faster turnaround time.

Limitations

There are several limitations to using next-generation sequencing. Next-generation sequencing provides information on a number of molecular aberrations. For many of the identified abnormalities, the clinical significance is currently unknown. Next-generation sequencing also requires sophisticated bioinformatics systems, fast data processing and large data storage capabilities, which can be costly. Although many institutions may have ability to purchase next-generation sequencing equipment, many lack the computational resources and staffing to analyze and clinically interpret the data.

Time and costs

The time to perform most next-generation sequencing methods and receive results has been greatly reduced. Starting from the day the laboratory receives the tumor specimen, it takes approximately 10 days for a physician to receive a whole-genome sequencing report.

Costs of sequencing the whole human genome have decreased significantly over the last decade. In 2006, the cost was approximately $20 million to $25 million. In 2016, the cost to sequence the human genome is generally less than $1,000.

Importance of Bioinformatics

The field of computer science called bioinformatics is used to analyze whole-genome sequencing data. This involves algorithm, pipeline and software development, and analysis, transfer and storage/database development of genomics data.

A typical whole-genome sequencing workflow contains the following steps:

1. quality control and data grooming;
2. genome assembly and/or variant calling; and
3. post-assembly analysis.

The volume of data that is produced from next-generation sequencing platforms is massive. Data collected pertains not only to the DNA sequencing results but also on the sequencing performance to assist with detection of errors or repetitive sequencing. This presents data management and storage issues. Additionally, special software and fast computing systems are required to process the immense data. Specialized, trained bioinformaticists are essential to the analysis of data generated by next-generation sequencing, as well as the continued success and growth of precision medicine.

Introduction to Targeted Gene Sequencing

Targeted gene sequencing panels are useful tools for analyzing specific mutations in a given sample. Focused panels contain a select set of genes or gene regions that have known or suspected associations with the disease or phenotype under study. Gene panels can be purchased with preselected content or custom designed to include genomic regions of interest.

Next-generation sequencing (NGS) offers the scalability, speed, and resolution to evaluate targeted genes of interest. Multiple genes can be assessed across many samples in parallel, saving time and reducing costs associated with running multiple separate assays. Targeted gene sequencing also produces a smaller, more manageable data set compared to broader approaches such as whole-genome sequencing, making analysis easier.

Targeted sequencing offers unique insights into specific regions of interest in the genome. It is a powerful application for investigating a variety of disease areas, such as oncology, inherited diseases, immunology and infectious diseases. This application allows targeting of specific genes, coding regions, even segments of chromosomes with precision and efficiency.

Targeted sequencing is more cost-effective than whole genome sequencing (WGS). It also enables deeper analysis of results than WGS and other survey approaches. In addition, it allows for deeper sequencing, and the depth of coverage helps in avoiding false interpretations of sequencing data. Because of this sensitivity, targeted sequencing provides tremendous advantage in variant calling in cancer research, identification of disease-associated mutations, single gene disorders and in gene expression studies. Targeted sequencing of specific regions also enables the discovery of causative genes for rare diseases.

The focused approach of targeted sequencing provides the possibility of its use in targeted therapy applications and in personalized medicine efforts. For example, targeted resequencing of the polymorphic human leucocyte antigen (HLA) gene helps in HLA typing, which is crucial for matching in hematopoietic stem cell or solid organ transplantation.

Advantages of Targeted Gene Sequencing

- Sequences key genes or regions of interest to high depth (500–1000× or higher), allowing identification of rare variants
- Provides cost-effective findings for studies of disease-related genes
- Delivers accurate, easy-to-interpret results, identifying variants at low allele frequencies (down to 5%)
- Enables confident identification of causative novel or inherited mutations in a single assay

- Allows detection and quantification of rare and low-frequency variants
- Enables higher coverage, deeper sequencing and straightforward data analysis
- Provides cost effectiveness, time and resource savings, and precision
- Offers a more manageable dataset for subsequent bioinformatics analysis
- Allows more samples to be analyzed than whole genome sequencing
-

Predesigned Targeted Gene Panels

Predesigned panels contain important genes or gene regions associated with a disease or phenotype, selected from publications and expert guidance. By focusing on the genes most likely to be involved, these panels conserve resources and minimize data analysis considerations. Predesigned panels are available for research on various diseases, such as cancer, inherited disorders, cardiac conditions, and autism.

Custom Targeted Gene Sequencing Solutions

With custom designs, researchers can target regions of the genome relevant to their specific research interests. Custom targeted sequencing is ideal for examining genes in specific pathways, or for follow-up experiments from genome-wide association studies or whole-genome sequencing.

Illumina supports two methods for targeted sequencing: target enrichment and amplicon generation.

- **Target enrichment**: Regions of interest are captured by hybridization to biotinylated probes and then isolated by magnetic pulldown. Target enrichment captures 20 kb–62 Mb regions, depending on the experimental design.
- **Amplicon sequencing**: Regions of interest are amplified and purified using highly multiplexed oligo pools. This method allows researchers to sequence a few genes to hundreds of genes in a single run, depending on the library preparation kit used.

DesignStudio Software: An easy-to-use online software tool that provides dynamic feedback to optimize probe designs.

AmpliSeq for Illumina Custom Panels: Create custom targeted sequencing panels optimized for content of interest.

Disease to Gene Finder Tool: Search by disease and find a ranked list of associated genes to help you design your custom panel or microarray.

Nextera Rapid Capture Custom Enrichment Kit: Custom assay for enriching genomic regions of interest, with add-on functionality.

How does targeted sequencing work?

The sample preparation workflow for targeted sequencing requires an additional step of target enrichment. It uses user-defined probe sets to enrich specific genomic regions of interest, thus causing only that region to be sequenced. The two methods for target enrichment are based on hybridization or amplification. While hybridization-based method uses probes to capture regions of interest, amplicon-based method uses PCR for target enrichment.

The hybridization-based target enrichment method

In the hybrid capture method, the process starts as a standard library preparation workflow. DNA is fragmented by shearing or using enzymes. Then adapters specific for the sequencing platform are added. Next, they are incubated with pools of biotinylated oligonucleotide probes designed to target specific regions of interest within a DNA fragment library. Finally, streptavidin-coated magnetic beads are used to attract the biotinylated probe/target hybrids. This method results in a sequencing-ready library that is highly enriched for the targeted DNA.

The PCR-based target enrichment method

In PCR-based methods, both uniplex and multiplex PCR reactions can be used. In multiplex PCR several primers targeted toward different target genes are used to generate multiple amplicons in a single reaction. After amplification, a normalization step is carried out for normalizing the concentration of the multiple PCR products. Then the pooled products are sequenced. While this method is efficient and easy to use, it is not ideal for targeting large genomic regions due to the cost of reagents for multiple reactions. Failure of targets to amplify and PCR bias are other drawbacks associated with PCR-based enrichment methods.

Our Roche offering

Robust target enrichment and construction of libraries with maximum molecular complexity and minimal bias is critical for targeted sequencing applications. Roche offers performance-optimized hybridization-based probes, both as fixed designs and custom panels.

In addition, Roche also offers an integrated approach to sample preparation using its validated sample preparation solutions encompassing all the steps required (from sample collection to quantification) to convert a sample to a sequencing-ready library.

6) Base Quality Score Recalibration

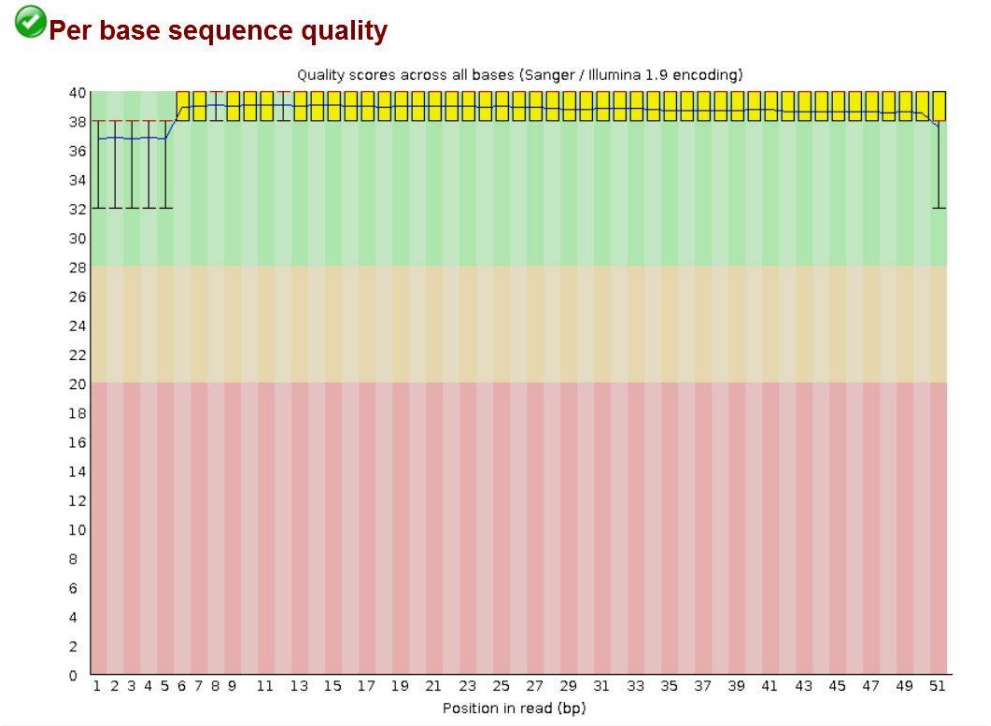
Variant calling algorithms rely heavily on the quality score assigned to the individual base calls in each sequence read. This is because the quality score tells us how much we can trust that particular observation to inform us about the biological truth of the site where that base

aligns. If we have a basecall that has a low quality score, that means we're not sure we actually read that A correctly, and it could actually be something else. So we won't trust it as much as other base calls that have higher qualities. In other words we use that score to weigh the evidence that we have for or against a variant allele existing at a particular site. [https://software.broadinstitute.org/gatk/best-practices/bp_3step.php?case=GermShortWGS]

Refresher: What are quality scores?

- Per-base estimates of error emitted by the sequencer
- Expresses the level of confidence for each base called
- Use standard Pred scores: Q20 is a general cutoff for high quality and represents 99% certainty that a base was called correctly
- 99% certainty means 1 out of 100 expected to be wrong. Let's consider a small dataset of 1M reads with a read length of 50, this means 50M bases. With 99% confidence, this means 50,000 possible erroneous bases.

The image below shows an example of average quality score at east position in the read, for all reads in a library (output from FastQC)



The image below shows individual quality scores (blue bars) for each position in a single read. The horizontal blue line represents the Q20 phred score value.

Quality scores emitted by sequencing machines are biased and inaccurate

Unfortunately the scores produced by the machines are subject to various sources of systematic technical error, leading to over- or under-estimated base quality scores in the data. Some of these errors are due to the physics or the chemistry of how the sequencing reaction works, and some are probably due to manufacturing flaws in the equipment. Base quality score recalibration (BQSR) is a process in which we apply machine learning to model these errors empirically and adjust the quality scores accordingly. This allows us to get more accurate base qualities, which in turn improves the accuracy of our variant calls. [https://software.broadinstitute.org/gatk/best-practices/bp_3step.php?case=GermShortWGS]

How BQSR works

1. You provide GATK Base Recalibrator with a set of known variants.
2. GATK Base Recalibrator analyzes all reads looking for mismatches between the read and reference, skipping those positions which are included in the set of known variants (from step 1).
3. GATK Base Recalibrator computes statistics on the mismatches (identified in step 2) based on the reported quality score, the position in the read, the sequencing context (ex: preceding and current nucleotide).
4. Based on the statistics computed in step 3, an empirical quality score is assigned to each mismatch, overwriting the original reported quality score.

   As an example, pre-calibration a file could contain only reported Q25 bases, which seems good. However, it may be that these bases actually mismatch the reference at a 1 in 100 rate, so are actually Q20. These higher-than-empirical quality scores provide false confidence in the base calls. Moreover, as is common with sequencing-by-synthesis machines, base mismatches with the reference occur at the end of the reads more frequently than at the beginning. Also, mismatches are strongly associated with sequencing context, in that the dinucleotide AC is often much lower quality than TG.

   [http://gatkforums.broadinstitute.org/gatk/discussion/44/base-quality-score-recalibration-bqsr]

2.1. Methods of targeted sequencing

Targeted sequencing comes in two main forms, amplicon or capture-based (Fig. 1). Amplicon-based enrichment utilises specifically designed primers to amplify only the regions of interest prior to library preparation [24]. Alternatively, in capture-based approaches, the DNA is fragmented and targeted regions are enriched via hybridization oligonucleotide bait sequences attached to biotinylated probes, allowing for isolation from the remaining genetic material [24,25]. Amplicon-based enrichment is the cheaper of the two technologies and shows a greater number of on target reads; however, the coverage of these regions is more uniform with hybrid sequencing [24,26]. Some commercially available amplicon platforms attempt to address the coverage issues by using specific primers that are able to amplify overlapping fragments in a single PCR reaction [27]. Amplicon based sequencing requires much less starting material than hybrid-capture, making it ideal if there is little DNA available for TS.

Hybrid-capture has been shown to produce fewer PCR duplicates than amplicon enrichment (<40% and up to ~80%, respectively) [24]. These duplicates are also more trivial to remove computationally, as the random shearing of the DNA in hybrid-capture platforms reduces the likelihood of two unique fragments aligning to the same genomic coordinates compared with the identical amplicons generated by amplicon enrichment platforms. This makes hybrid-capture especially useful for samples where these PCR artefacts are more likely to occur, such as FFPE and ctDNA samples. Further, certain regions of the genome make primer design for amplicon enrichment difficult (e.g. regions with a high number of repeated sequences). The long bait sequences used in hybrid-capture, however, allow a greater level of specificity in region selection. Overall hybrid-capture based platforms provide more accurate and uniform target selection, whilst amplicon-based platforms are often used in small scale experiments where sample quantity or cost are a factor.

2.2. Platforms for targeted sequencing

There are several commercially available platforms for these two approaches. Many of these platforms are also used for WES. An outline of these platforms is shown in Table 3. Despite the differences between the various platforms, they have been shown to lead to relatively concordant variant calling [24].

2.3. Use and design of panels for targeted sequencing

2.3.1. Targeted panel construction

The term targeted panel is used here to refer to the collection of genomic coordinates that are of interest to the user. An important difference between WES panels and targeted panels, is that TS is not constrained to canonical gene targets and can target other regions, such as promoters [28] or breakpoints [29]. There are commercially available targeted gene panels, usually designed for research [30,31] or clinical purposes [32,33]. They are designed to amplify genomic regions that are known to be of interest within cancer, or specific cancer subtypes. Using these panels greatly speeds up the process of the sequencing as they have already been designed, tested and validated. Commonly, however, users design their own customised panels dependent on their research questions, although thorough target validation of these panels is needed before use. Customised panels are often generated by a thorough review of the current literature and cross referencing publicly available cancer mutation resources such as TCGA, ICGC, CbioPortal, and Catalogue of Somatic Mutations in Cancer (COSMIC) (http://cancer.sanger.ac.uk) databases [34–38], selecting genes that are frequently mutated, and targets that have been functionally validated in that cancer. In many cancer studies, an initial discovery cohort has been initially profiled with WGS or WES to the identify significantly mutated genes (via algorithms like MutSigCV [39], dNdScv [40], oncodriveFM [41]). These genes are then selected for TS with higher depth in the validation cohort(s) to establish their validity and frequencies [42– 45]. Examples of the applications of these panels are included in the next section.

2.3.2. Applications of targeted gene panels in cancer studies There are a large body of clinical studies that utilise genomic TS for research on clinical samples. Some recent examples have been listed in Table 4 [17,43–50], with targeted panels ranging from as few as 25 genes [44] to 122 genes [49]. These studies illustrate that a wide range of TS platforms, sequencing depths, data processing and variant calling methods were used.

3. Guidance for analysis of targeted genomic sequencing

In this section we provide detailed guidance for the analysis of TS, from initial quality control (QC) and data pre-processing, to variant calling, annotation and filtering (Fig. 2). Commonly used methods and software in each step and important parameters/filters are discussed, aiming to provide readers a comprehensive overview of the whole analytical process from raw reads to highconfidence annotated calls. We further focus on PCR duplication marking/removal and variant filtering in greater depth, as these are crucial steps to ensure the best quality variant calls. Key steps of TS data analysis and commonly used software are listed in Table 5.

3.1. Quality control and data pre-processing

3.1.1. QC and alignment

The first step of all NGS pipelines is to assess the quality of the sequenced reads, using FastQC (http://www.bioinformatics.bbsrc. ac.uk/projects/fastqc). It summarises and

visualises base quality score for every base pair sequenced, which allows users to have an overview of the read quality and decide whether a trimming step is needed, especially at the 30 end where the base quality is often lower. FastQC also produces summarised information of adapter fragment contamination and GC content within all reads. This analysis determines whether adapter fragments have been incorporated into the reads and need to removed using software such as CutAdapt [51]. The GC content of the reads is useful to indicate whether the sample is contaminated with DNA from another organism, as this would likely lead to a secondary peak due to the different GC content of that genome [63]. Next, raw or trimmed reads are aligned to the reference genome to generate Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) files for each sample. Commonly used aligners include the Burrows-Wheeler Aligner (BWA) [53] and Bowtie2 [54]. Ion TorrentTM also have their own customised aligner specifically for working on data generated from their platform. Within alignment files the mapping quality score (i.e., the likelihood of a read mapping to multiple locations in the genome) is recorded for each read, in addition to their mapped coordinates. It should be noted that the experimental and web-lab quality of TS experiments is also a key determinant of the sequencing data quality, such as how fragmented the DNA is, and the amount of input DNA. Low quantity of input DNA will require more PCR cycles, leading to a high level of PCR duplicates and limiting the achievable depth of coverage of the experiment. Monitoring the experimental quality of TS is always part of good laboratory practice, ensuring the highest quality of sequencing data in the downstream analyses. It is also important to check for germline/tumour mix-ups and contamination whilst running the pipeline. Whilst these errors are very difficult to determine from the FASTQ files alone, they may become more apparent in the later analytic stages, such as variant calling and VAFs, e.g. a large number of variants called in the germline that are absent in the tumour sample.

3.1.2. Assessment of off-target reads

Various QC steps should always take place to ensure the best quality of TS data. As TS focuses on regions of interest in the design panel, we expect the majority of reads generated should come from targeted regions, however, off-target reads are a common occurrence. After alignment, the percentage of reads that cover targeted regions can be assessed using software such as bedtools [52], and the GATK coverage module. A high proportion of off-target reads may indicate that the TS experiment has failed, or the targeted regions contain too many repeat sequences. This could be possibly adjusted by making the capture or library preparation process more efficient, e.g., adjust input DNA to beads ratio, and wash more stringently. With a large panel of hundreds of targeted genes, roughly >70% of the reads aligning to the targeted regions is a positive indicator of a good quality TS data set [26].

3.1.3. Marking and removal of PCR duplicates

PCR duplicates are sequence reads that align to the same genomic coordinates and typically arise during PCR steps in the library preparation. The duplication rate tends to be much higher for fragmented DNA of low quality, e.g. FFPE and ctDNA, reaching ~50–60% for some cases, while for FF DNA, this rate is usually less than 20%. These PCR duplicates need to be marked and removed before any downstream analysis, as including them will lead to

overestimation of coverage in targeted regions, and more importantly result in incorrect allele frequency estimation. A number of software are used to search for PCR duplicates within aligned NGS data. A commonly used program is the MarkDuplicates function within Picard Tools (http://broadinstitute.github.io/picard/). This tool looks for reads with the same start and end coordinates and then add tags to the bam files that mark these reads as duplicates. Another tool, SAMtools rmdup, simply outright removes the duplicate reads retaining the read with the highest mapping quality [55]. However, these software based attempts cannot discriminate between two unique reads that happen to align in the same position by chance and actual duplicates [64]. There are additional molecular techniques, such as Unique Molecular Identifiers or Molecular Barcodes (MBC), available to ensure only unique reads are measured in the downstream analysis. These are exemplified by the Nonacus Cell3TM Target, Agilent HaloplexHS and SureSelectXT platforms.

3.1.4. Realignment, base score recalibration and estimation of sequencing coverage

Next, filtered alignments are further processed to improve the alignment quality, including local realignment around indels and base quality score recalibration using GATK. The step of local realignment is to improve the alignment quality for bases around known and suspected indel positions to reduced false positive calls. Base score recalibration is carried out to recalculate base quality scores for all sequenced reads based on known polymorphisms (e.g., SNPs from 1000G Project). The base and mapping quality scores are used to filter reads during variant calling and the fine-tuning that occurs in this step is important to ensure only high-confidence variants are called. Base coverage information is another important parameter to assess the overall quality of TS data. Using recalibrated BAM files, one can further calculate the coverage/depth for bases within the targeted regions, using Bedtools or GATK coverage. Depending on the quality of DNA and total number of reads generated, several hundred times depth per base is often expected, although some regions may have much higher coverage or targeted rates than others. However, for ultra-deep sequencing, the depth of tens of thousands of reads is often required to detect very low frequency clones.

3.2. Variant calling

Once all TS pre-processing steps are completed, these highquality alignment data are ready for variant calling. Variant calling is the process of comparing the aligned reads to a reference genome or matched normal DNA sequences to identify base pair variations. Here we describe the procedure for samples with matched normal and without matched normal separately. We then focus on variant calling parameters and filters which can be tuned accordingly to achieve the best outcome.

Variant calling parameters and filtering

A set of important parameters need to be considered for variant calling and filtering for high-quality calls. These include, Number of total reads: this parameter can be used to ensure there is sufficient coverage over the position for variants to be called. Often a minimum of 20-30x depth is required for TS [71–75]. Number of variant supporting reads: this parameter should be set in order to limit variants with very few supporting reads being considered. The value

can be tuned based on the average coverage of the samples. Usually the minimum value ranges from 4 to 10 reads [26,47,76]. Minimum base and mapping quality score: Setting a threshold for base and mapping quality scores stops poorly sequenced or aligned reads from being considered in the variant calling. F. Bewicke-Copley et al. / Computational and Structural Biotechnology Journal 17 (2019) 1348–1359 1355 The default minimum values of many programmes are set as 20–30 as these correspond to an accuracy of 99% and 99.9% respectively

Minimum allele frequency for called variants:

Like the number of variant supporting reads, this can be used to eliminate variant positions with low levels of support. Often, a relatively low threshold (e.g., 3% with a depth of 200x) is initially used to include most of the variants, and further filtering and refinement are performed via testing a range of threshold values to choose the best cutoff value for VAF. For FFPE samples, the final threshold is set as at least 10% or even 20% across many studies [77,78]. For FF samples this threshold can be much lower depending on overall sequencing depth [46,50]. One should note that the tumour purity of clinical samples is often highly heterogeneous. Thus, filtering simply based on an observed VAF cutoff may not provide the most accurate way to include high-quality or exclude low-quality calls. One way to overcome this is to further adjust VAF values based on the estimates of tumour purity of clinical samples, and apply the threshold on these adjusted VAFs to filter calls for the downstream analyses. When an accurate measurement of tumour purity is not available, VAFs of mutations in many known clonal driver genes (e.g., KRAS and TP53 for many solid tumours) could be used to derive a rough estimate.

Additional parameters also include:

Strandedness of variant supporting reads:

If a variant occurs within a sample, paired sequencing should show evidence of this variant on both strands. Therefore if the majority of the reads for a variant occur on only one strand (i.e., strand bias), it could suggest that variant reads are artefacts [58,76]. In many programmes, at least one supporting read is required to be present on each strand for the called variants. In VarScan2, it is possible to require that a maximum of 90% of all reads (across reference and alternative alleles) are found on one strand, meaning positions that have a strand bias will be ignored.

Significance score for a statistical test:

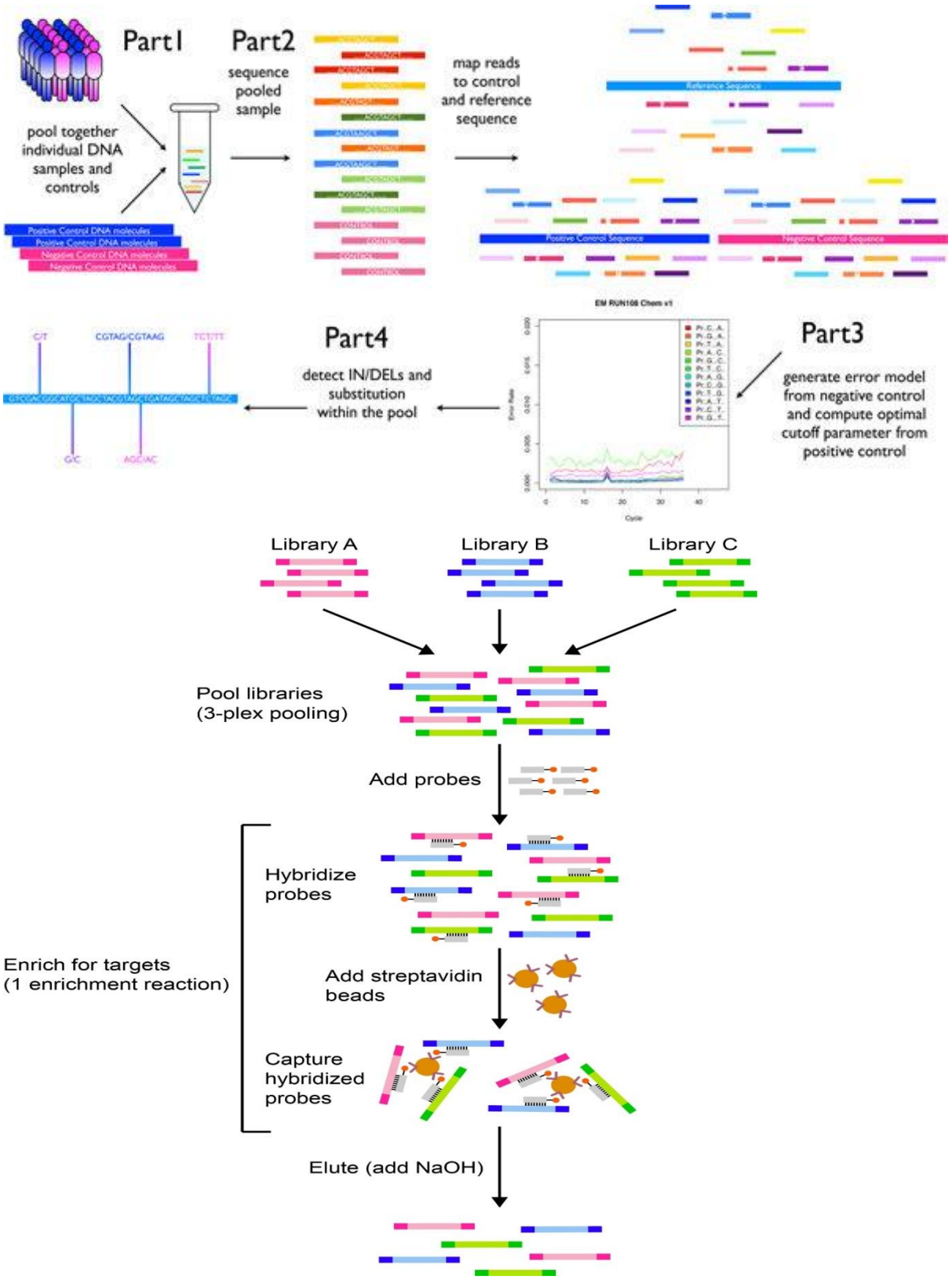
Many variant callers will calculate a statistical evaluation of the likelihood of a variant differing from the reference allele [47,76]. VarScan2 for example provides the user with a p value for a Fisher’s Exact Test on the observed and expected variant reads. This can be used to further eliminate low-quality calls.

3.3. Annotation and further filtration of variants. Following variant calling, the next step is to annotate the variants in relation to genes (e.g., within or outside a gene), codon and amino acid positions, and classify types of variants, such as nonsense, missense, exonic deletions

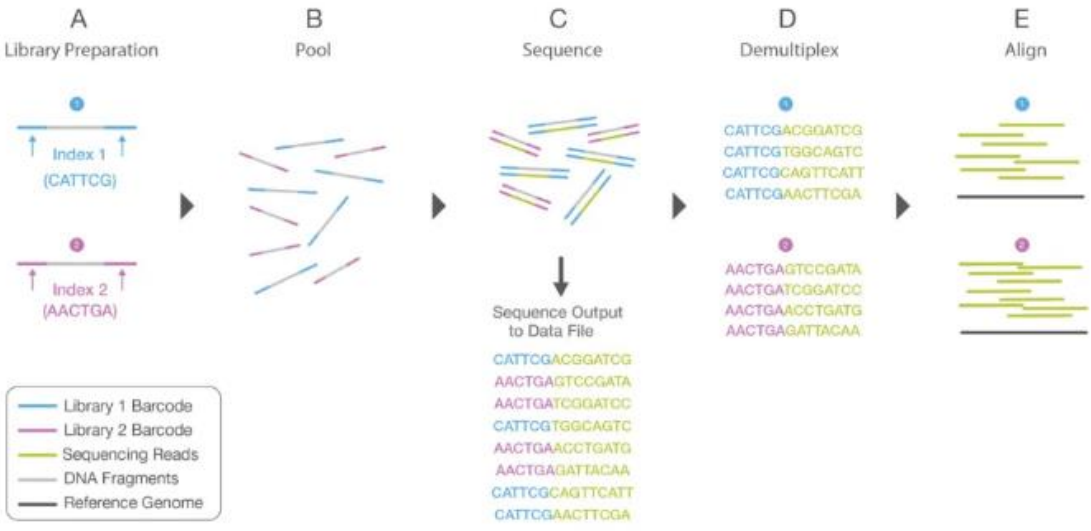
and synonymous variants. This allows for greater understanding of their functional consequences on genes they relate to.

Pooled sequencing

- Cost reduced
- Producing tens of thousands of genomes, or so-called ‘ will revolutionize the study of population diversity and help us to genetic basis of health and disease better
- The main challenge exists in individually amplifying and creating sequencing libraries for thousands of samples. To efficiently use the capacity of sequencer and reduce the cost of sequencing library construction for large-scale sequencing, multiple individuals could be pooled together and sequenced, called pooled sequencing (pool-seq).
- Pool-seq could provide a cost-effective alternative to sequencing individuals separately, since pool-seq uses a single library for the entire sample, whereas sequencing of individuals requires a separate library to be prepared for each sample
- Pool-seq could save tremendously on sample prepara-tions, especially for targeted sequencing projects, sincethe cost for target capturing is proportional to the numberof samples (i.e., number of individuals without pooling vs. number of pools in pool-seq)
- multiple populations or generations


Part1
Part2
sequence pooled sample
pool together individual DNA samples and controls
map reads to control and reference sequence
Reference Sequence
Positive Control Sequence
Negative Control Sequence
Part3
generate error model from negative control and compute optimal cutoff parameter from positive control
Part4
detect IN/DELs and substitution within the pool
Library A
Library B
Library C
Pool libraries (3-plex pooling)
Add probes
Hybridize probes
Add streptavidin beads
Capture hybridized probes
Enrich for targets (1 enrichment reaction)
Elute (add NaOH)

- The main limitation of the naive pool-seq strategy is its inability to obtain the information for each individual sample participated in the pool. However, multiplexing using sequencing barcodes could overcome the drawback, where the DNA in each sample is cut into short fragments suitable for sequencing and ligated with a short, sample-specific DNA sequence i.e. barcode [8]. After sequencing, reads belonging to each individual could be assigned precisely based on the barcode signature.
- effective in SNP discovery and could provide more accurate allele frequency estimates at a lower cost than sequencing of individuals, even when taking sequencing errors into account
- Mainly applied in genome-wide association studies (GWAS), population genetics, reverse ecology, genome evolution studies
- In 2009, Pattersonet al. proposed the concept of the combination between combinatorics and pooled sequencing, defined as combinatorial pooled sequen-cing which allows the sequencing results to be decoded to identify the reads belonging to samples that are unique or rare among the population without barcodes. Using ideas from a branch of mathematics called combinatorics, thousands of samples are pooled and sequenced at the same time in the combinatorial pooled sequencing.
- in many applications, such as identifying rare variants carriers and rare haplotype carriers, assembling complex genome, single individual haplotyping, sequencing of multiple viral samples.



- DNA barcoding, DNA Sudoku, comparing the estimated allele frequencies between cases and controls without actually inferring individual genotypes.
- The savings on cost and time come from two sources. The first is that estimating the allele frequency requires much less depth of coverage per individual than that

required for calling the genotype of each individual. The second is the reduced efforts in library preparation for a large number of DNA samples.

Advantages

- accurate estimate of the allele frequency,
- potential to detect rare variants

Prerequisites for the pooling of customer libraries are:

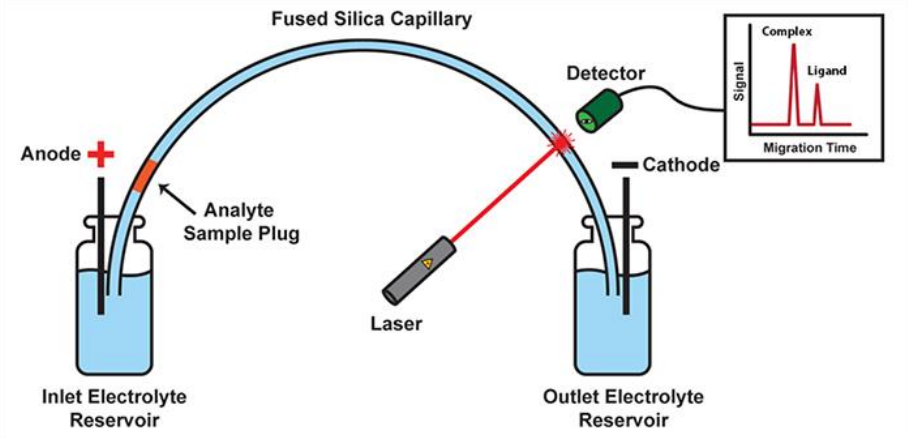
- all libraries were generated using the same protocol and are PCR amplified
- the **library fragment sizes have to be similar for all libraries**(and within Illumina specs) as demonstrated by Bioanalyzer traces (or gel images if correct balancing is not that critical)
- have uniquely indexed adapters
- all libraries have DNA concentrations in the same range
- PCR-amplified libraries can be quantified based on fluorometric measurements (e.g. Qubit), but PCR-free libraries are best quantified by qPCR.

Resequencing

- Resequencing of candidate genes or other genomic regions of interest in patients and controls is a key step in detection of mutations associated with various congenital diseases.
- Resequencing techniques can be divided into those which test for known mutations (genotyping) and those which scan for any mutation in a given target region (variation analysis).
- Typical mutations being tested are substitution (SNP), insertion and deletion mutations.

Electrophoresis-based resequencing.

- One of most advanced resequencing techniques based on electrophoresis is capillary electrophoresis.
- VariantSEQr™ system created by Applied Biosystems successfully integrates capillary electrophoresis, PCR and state-of-the-art automated data analysis techniques for quick and accurate resequencing of specific human genes.
- **Capillary electrophoresis** is a separation technique in which charged species are separated, based on charge and size, by their different rates of migration in an electric field. The **capillary** is made of negatively charged fused silica inner wall which forms an electrical double layer with cations in the running buffer.



Array-based high-throughput resequencing strategies.

I. The gain of hybridization signal approach.

- In the gain of hybridization signal approach, relative hybridization to allele-specific probes complementary to each of the four possible nucleotides at interrogated nucleotide position is used for genotype analysis.
- Disadvantage: analysis requires large amount of carefully designed probes, for example, interrogating both target strands of length N for all possible insertions of length X requires $2(4^X)N$ probes.

II. The loss of hybridization signal approach

- In the loss of hybridization signal approach, decreased hybridization of red-labeled test target relative to green-labeled reference target to perfect match probes interrogating the area of interest indicates the presence of a sequence change.
- Disadvantage: the mutation cannot be discerned; the identity of the sequence change must be established by subsequent dideoxysequencing of the region surrounding the loss of signal signature.

III. The minisequencing approach.

- In the minisequencing approach unlabeled target is hybridized to perfect match probes (attached through a 5' linkage to the array to leave an exposed 3'-OH group) interrogating the nucleotide position of interest. Fluorescently tagged ddNTPs are used in subsequent enzymatic primer extension reactions to extend the hybridized primers. The identity of the extended ddNTP is used in sequence analysis.
- Disadvantage: designing and validating primers can be a long, tedious process that often leads to experimental delays and defective PCR products; data analysis requires not only a high level of expertise but also substantial time commitment.

Whole Genome Resequencing

- The goal of a WGR experiment is usually to identify the differences between the genome of specific individuals and that of a so called, reference genome.
- By comparing the sequenced genomes to the reference, a catalog of mutations specific to each sequenced individual is obtained, usually single nucleotide polymorphisms (SNPs) and insertions- deletions (indels), which can provide an extremely valuable insight into the genetic background of the individuals. Often this is associated with specific phenotypes based on which the individual are selected. Additionally, with specifically planned experiments, large rearrangements (e.g., translocations, inversions, large copy number variations) can be also pinpointed through WGR.

Whole Exome sequencing

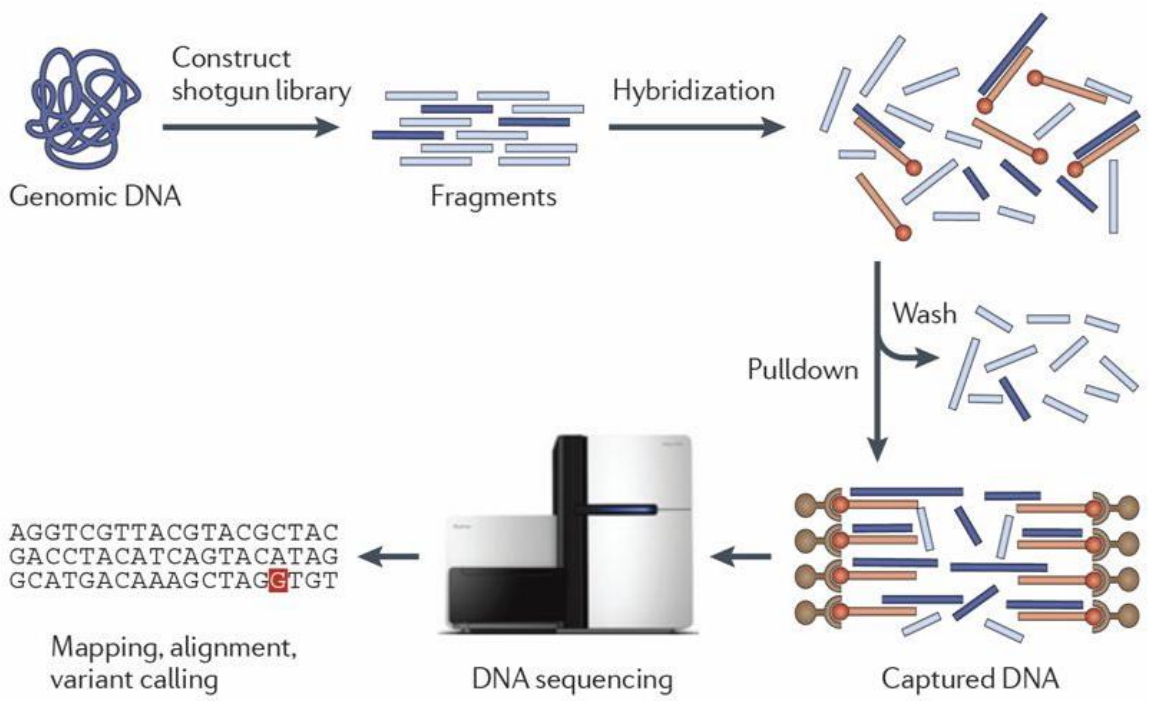
- Whole-exome sequencing is a widely used next-generation sequencing (NGS) method that involves sequencing the protein-coding regions of the genome. The human exome represents less than 2% of the genome, but contains ~85% of known disease-related variants,[1] making this method a cost-effective alternative to whole-genome sequencing.
- Exome sequencing using exome enrichment can efficiently identify coding variants across a broad range of applications, including population genetics, genetic disease, and cancer studies.

Advantages of Exome Sequencing

- Identifies variants across a wide range of applications
- Achieves comprehensive coverage of coding regions
- Provides a cost-effective alternative to whole-genome sequencing (4–5 Gb of sequencing per exome compared to ~90 Gb per whole human genome)
- Produces a smaller, more manageable data set for faster, easier data analysis compared to whole-genome approaches

Exome sequencing detects variants in coding exons, with the capability to expand targeted content to include untranslated regions (UTRs) and microRNA for a more comprehensive view of gene regulation. DNA libraries can be prepared in as little as 1 day and require only 4–5 Gb of sequencing per exome.

Exome Sequencing



Bamshad, MJ., et al. *Nat. Rev. Genet.* (2011) 12:745-755.

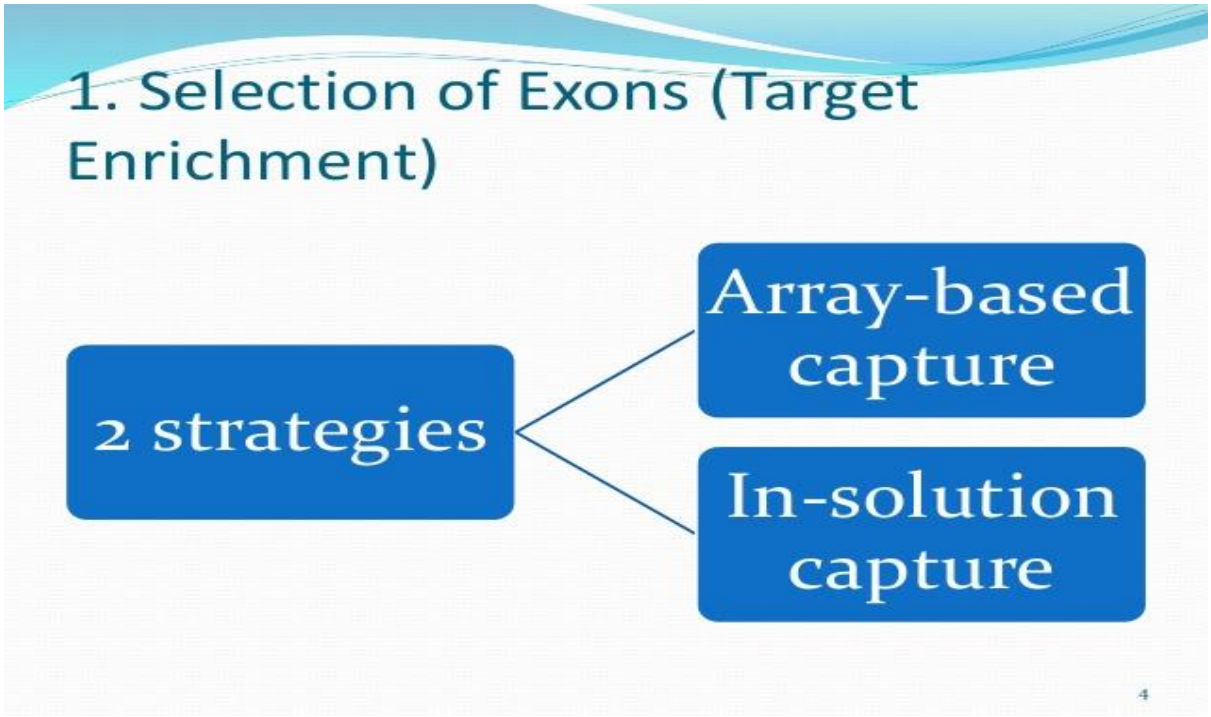
Methodology

2 steps:

1. Select only the subset of DNA that encodes proteins (exons)

2. Sequencing of the exonic DNA using DNA sequencing technology

3

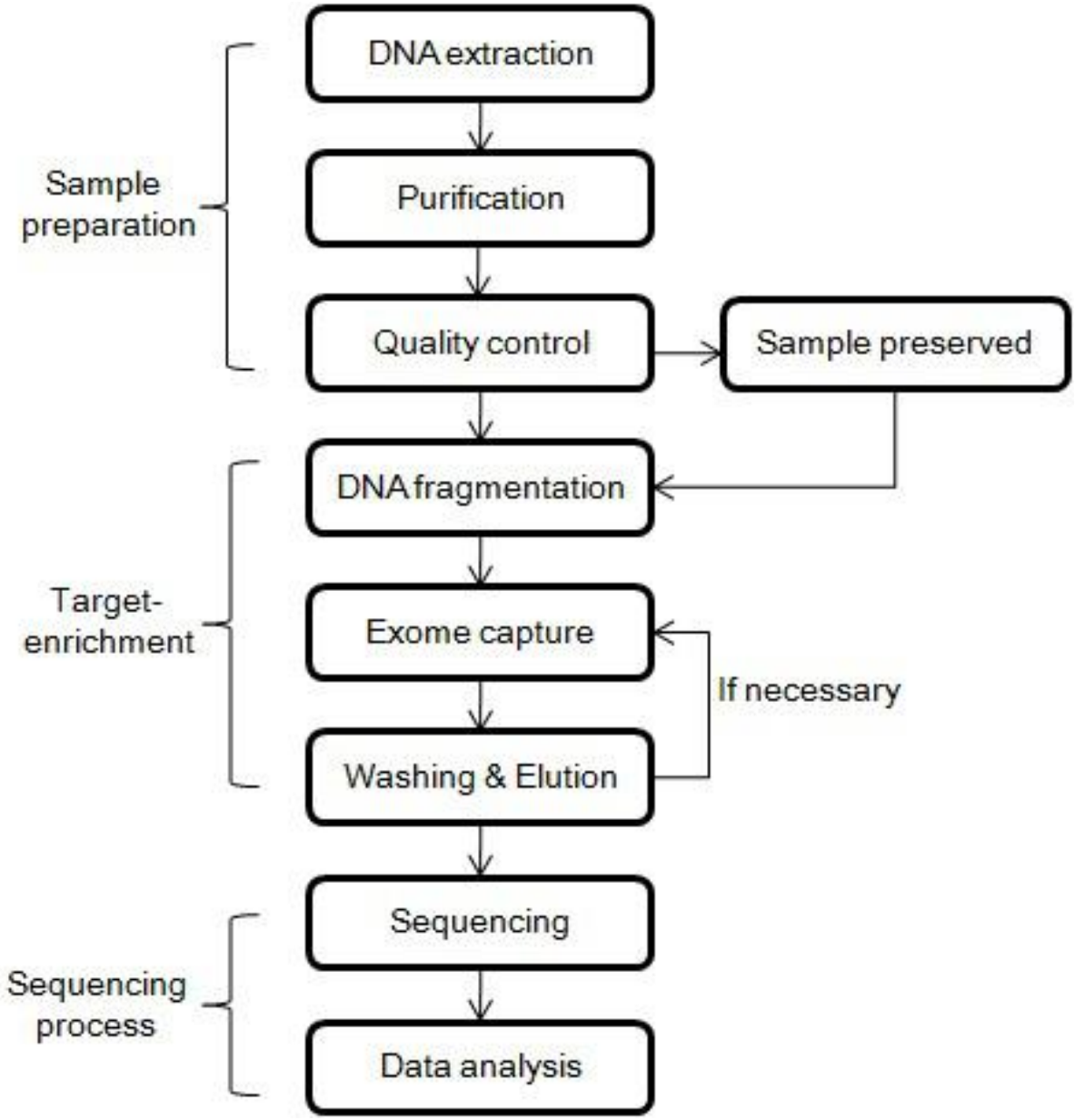


Array-based capture

- Microarrays contain single-stranded oligonucleotides with sequences from the human genome to tile the region of interest fixed to the surface. Genomic DNA is sheared to form double-stranded fragments. The fragments undergo end-repair to produce blunt ends and adaptors with universal priming sequences are added. These fragments are hybridized to oligos on the microarray. Unhybridized fragments are washed away and the desired fragments are eluted. The fragments are then amplified using PCR

In-solution capture

- To capture genomic regions of interest using in-solution capture, a pool of custom oligonucleotides (probes) is synthesized and hybridized in solution to a fragmented genomic DNA sample. The probes (labeled with beads) selectively hybridize to the genomic regions of interest after which the beads (now including the DNA fragments of interest) can be pulled down and washed to clear excess material. The beads are then removed and the genomic fragments can be sequenced allowing for selective DNA sequencing of genomic regions (e.g., exons) of interest.
- Magnetic bead is a kind of magnetic nanoparticles which contain functional chemical components to combine target substances. In this case, magnetic beads which could bind exome are used.



Applications

- Rare variant mapping in complex disorders
- Discovery of Mendelian disorders
- Clinical Diagnostics

Variants in an exome

- An exome analysis pipeline identifies about 20000-25000 variants (pass filter variants)
- On average an exome contain
  - ~10 000 silent
  - ~10 000 missense *(alter protein sequence)*
  - ~100 nonsense *(truncate protein sequence)*
  - ~40 splice *(affect splicing patterns)*
- Bottleneck of current methods is to identify disease-causing variant(s)

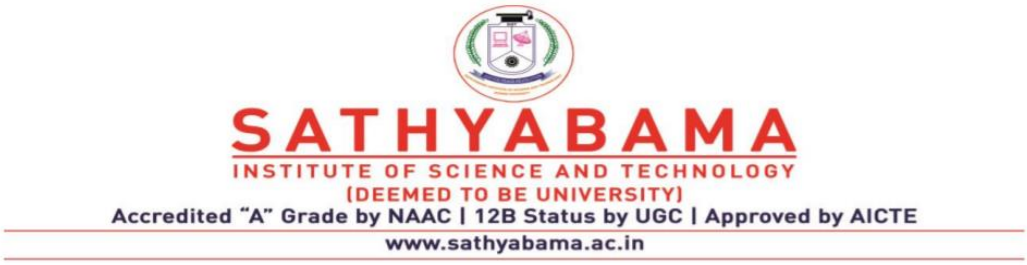
Exome sequencing

Benefits

- **Compared to whole-genome sequencing**
  - Low cost, Analysis is fast, data storage
  - Majority of functional variants in lower cost
- **Compared to traditional methods**
  - All coding variants in one experiment
  - Higher success rate
  - Identifying rare, de novo, and novel genes
  - Expanding the phenotypic spectrum
  - Patients with more than one Mendelian disorder
  - Modifier genes
  - Insight for research

Challenges

- **Compared to whole-genome sequencing**
  - Exome capture efficiency.
  - Genome-wide variants
- **Compared to traditional methods**
  - Genes / exons in highly homologous regions
  - CNV and other structural variant detection
  - Triplet repeat expansions are not detected
  - Bioinformatics pipelines Incidental findings
  - Ethical issues



SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – III - NEXT GENERATION SEQUENCING – SBI1606

RNA Sequencing

Introduction

The central dogma of molecular biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins (Crick 1958; Crick 1970). The ultimate expression of this genetic information modified by environmental factors characterizes the phenotype of an organism. The transcription of a subset of genes into complementary RNA molecules specifies a cell's identity and regulates the biological activities within the cell. Collectively defined as the transcriptome, these RNA molecules are essential for interpreting the functional elements of the genome and understanding development and disease. The transcriptome has a high degree of complexity and encompasses multiple types of coding and noncoding RNA species. Historically, RNA molecules were relegated as a simple intermediate between genes and proteins, as encapsulated in the central dogma of molecular biology. Therefore, messenger RNA (mRNA) molecules were the most frequently studied RNA species because they encoded proteins via the genetic code.

In addition to protein coding mRNA, there is a diverse group of noncoding RNA (ncRNA) molecules that are functional. Previously, most known ncRNAs fulfilled basic cellular functions, such as ribosomal RNAs and transfer RNAs involved in mRNA translation, small nuclear RNA (snRNAs) involved in splicing, and small nucleolar RNAs (snoRNAs) involved in the modification of rRNAs (Mattick and Makunin 2006). More recently, novel classes of RNA have been discovered, enhancing the repertoire of ncRNAs. For instance, one such class of ncRNAs is small noncoding RNAs, which include microRNA (miRNA) and piwi-interacting RNA (piRNA), both of which regulate gene expression at the posttranscriptional level (Stefani and Slack 2008). Another noteworthy class of ncRNAs is long noncoding RNAs (lncRNAs). As a functional class, lncRNAs were first described in mice during the largescale sequencing of cDNA libraries (Okazaki et al. 2002). A myriad of molecular functions have been discovered for lncRNAs, including chromatin remodeling, transcriptional control, and posttranscriptional processing, although the vast majority are not fully characterized (Guttman et al. 2009; Mercer et al. 2009; Wilusz et al. 2009).

Initial gene expression studies relied on low-throughput methods, such as northern blots and quantitative polymerase chain reaction (qPCR), that are limited to measuring single transcripts. Over the last two decades, methods have evolved to enable genome-wide quantification of gene expression, or better known as transcriptomics. The first transcriptomics studies were performed using hybridization-based microarray technologies, which provide a high-throughput option at relatively low cost (Schena et al. 1995). However, these methods have several limitations: the requirement for a priori knowledge of the sequences being interrogated; problematic cross-

hybridization artifacts in the analysis of highly similar sequences; and limited ability to accurately quantify lowly expressed and very highly expressed genes (Casneuf et al. 2007; Shendure 2008). In contrast to hybridizationbased methods, sequence-based approaches have been developed to elucidate the transcriptome by directly determining the transcript sequence. Initially, the generation of expressed sequence tag (EST) libraries by Sanger sequencing of complementary DNA (cDNA) was used in gene expression studies, but this approach is relatively low-throughput and not ideal for quantifying transcripts (Adams et al. 1991, 1995; Itoh et al. 1994).

To overcome these technical constraints, tag-based methods such as serial analysis of gene expression (SAGE) and cap analysis gene expression (CAGE) were developed to enable higher throughput and more precise quantification of expression levels. By quantifying the number of tagged sequences, which directly corresponded to the number of mRNA transcripts, these tag-based methods provide a distinct advantage over measuring analogstyle intensities as in array-based methods (Velculescu et al. 1995; Shiraki et al. 2003). However, these assays are insensitive to measuring expression levels of splice isoforms and cannot be used for novel gene discovery. In addition, the laborious cloning of sequence tags, the high cost of automated Sanger sequencing, and the requirement for large amounts of input RNA have greatly limited its use. The development of high-throughput next-generation sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis through the sequencing of complementary DNA (cDNA) (Wang et al. 2009). This method, termed RNA sequencing (RNA-Seq), has distinct advantages over previous approaches and has revolutionized our understanding of the complex and dynamic nature of the transcriptome. RNA-Seq provides a more detailed and quantitative view of gene expression, alternative splicing, and allele-specific expression. Recent advances in the RNA-Seq workflow, from sample preparation to sequencing platforms to bioinformatic data analysis, has enabled deep profiling of the transcriptome and the opportunity to elucidate different physiological and pathological conditions. In this article we will provide an introduction to RNA sequencing and analysis using nextgeneration sequencing methods and discusses how to apply these advances for more comprehensive and detailed transcriptome analyses.

Transcriptome Sequencing

The introduction of high-throughput next-generation sequencing (NGS) technologies revolutionized transcriptomics. This technological development eliminated many challenges posed by hybridization-based microarrays and Sanger sequencing-based approaches that were previously used for measuring gene expression. A typical RNA-Seq experiment consists of isolating RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on an NGS platform (Fig. 1). However, many experimental details, dependent on a researcher's objectives, should be considered before performing RNA-Seq. These include the use of biological and technical replicates, depth of sequencing, and desired coverage across the transcriptome. In some cases, these experimental options will have minimal impact on the

quality of the data. However, in many cases the researcher must carefully design the experiment, placing a priority on the balance between high-quality results and the time and monetary investment.

Isolation of RNA

The first step in transcriptome sequencing is the isolation of RNA from a biological sample. To ensure a successful RNA-Seq experiment, the RNA should be of sufficient quality to produce a library for sequencing. The quality of RNA is typically measured using an Agilent Bioanalyzer, which produces an RNA Integrity Number (RIN) between 1 and 10 with 10 being the highest quality samples showing the least degradation. The RIN estimates sample integrity using gel electrophoresis and analysis of the ratios of 28S to 18S ribosomal bands. Note that the RIN measures are based on mammalian organisms and certain species with abnormal ribosomal ratios (i.e., insects) may erroneously generate poor RIN numbers. Lowquality RNA (RIN < 6) can substantially affect the sequencing results (e.g., uneven gene coverage, 3′–5′ transcript bias, etc.) and lead to erroneous biological conclusions. Therefore, high-quality RNA is essential for successful RNA-Seq experiments. Unfortunately, highquality RNA samples may not be available in some cases, such as human autopsy samples or paraffin embedded tissues, and the effect of degraded RNA on the sequencing results should be carefully considered

Library Preparation Methods

Following RNA isolation, the next step in transcriptome sequencing is the creation of an RNA-Seq library, which can vary by the selection of RNA species and between NGS platforms. The construction of sequencing libraries principally involves isolating the desired RNA molecules, reverse-transcribing the RNA to cDNA, fragmenting or amplifying randomly primed cDNA molecules, and ligating sequencing adaptors. Within these basic steps, there are several choices in library construction and experimental design that must be carefully made depending on the specific needs of the researcher (Table 1). Additionally, the accuracy of detection for specific types of RNAs is largely dependent on the nature of the library construction. Although there are a few basic steps for preparing RNA-Seq libraries, each stage can be manipulated to enhance the detection of certain transcripts while limiting the ability to detect other transcripts.

Selection of RNA Species—

Before constructing RNA-Seq libraries, one must choose an appropriate library preparation protocol that will enrich or deplete a “total” RNA sample for particular RNA species. The total RNA pool includes ribosomal RNA (rRNA), precursor messenger RNA (pre-mRNA), mRNA, and various classes of noncoding RNA (ncRNA). In most cell types, the majority of RNA molecules are rRNA, typically accounting for over 95% of the total cellular RNA. If the rRNA transcripts are not removed before library construction, they will consume the bulk of the sequencing reads, reducing the overall depth of sequence coverage and thus limiting the detection of other less-abundant RNAs. Because the efficient removal of rRNA is critical for

successful transcriptome profiling, many protocols focus on enriching for mRNA molecules before library construction by selecting for polyadenylated (poly-A) RNAs. In this approach, the 3′ poly-A tail of mRNA molecules is targeted using poly-T oligos that are covalently attached to a given substrate (e.g., magnetic beads). Alternatively, researchers can selectively deplete rRNA using commercially available kits, such as RiboMinus (Life Technologies) or RiboZero (Epicentre). This latter method facilitates the accurate quantification of noncoding RNA species, which may be polyadenylated and thus excluded from poly-A libraries. Lastly, highly abundant RNA can be removed by denaturing and re-annealing double-stranded cDNA in the presence of duplex-specific nucleases that preferentially digest the most abundant species, which re-anneal as double-stranded molecules more rapidly than lessabundant molecules (Christodoulou et al. 2011). This method can also be used to remove other highly abundant mRNA transcripts in samples, such as hemoglobin in whole blood, immunoglobulins in mature B cells, and insulin in pancreatic beta cells.

A comprehensive understanding of the technical biases and limitations surrounding each methodological approach is essential for selecting the best method for library preparation. For example, poly-A libraries are the superior choice if one is solely interested in coding RNA molecules. Conversely, ribo-depletion libraries are a more appropriate choice for accurately quantifying noncoding RNA as well as pre-mRNA that has not been posttranscriptionally modified. Furthermore, moderate differences exist between ribodepletion protocols, such as the efficiency of rRNA removal and differential coverage of small genes, which should be investigated before selecting a method (Huang et al. 2011).

In addition to the selective depletion of specific RNA species, new approaches have been developed to selectively enrich for regions of interest. These approaches include methods employing PCR-based approaches, hybrid capture, in-solution capture, and molecular inversion probes (Querfurth et al. 2012). The hybridization-based in solution capture involves a set of biotinylated RNA baits transcribed from DNA template oligo libraries that contain sequences corresponding to particular genes of interest. The RNA baits are combined with the RNA-Seq library where they hybridize to RNA sequences that are complementary to the baits, and the bounded complexes are recovered using streptavidincoated beads. The resulting RNA-Seq library is now enriched for sequences corresponding to the baits and yet retains its gene expression information despite the removal of other RNA species (Levin et al. 2009). The approach enables researchers to reduce sequencing costs by sequencing selected regions in a greater number of samples.

Selection of Small RNA Species—

Complementing the library preparation protocols discussed above, more specific protocols have been developed to selectively target small RNA species, which are key regulators of gene expression. Small RNA species include microRNA (miRNA), small interfering RNA (siRNA), and piwi-interacting RNA (piRNA). Because small RNAs are lowly abundant, short in length

(15–30 nt), and lack polyadenylation, a separate strategy is often preferred to profile these RNA species (Morin et al. 2010). Similar to total RNA isolation, commercially available extraction kits have been developed to isolate small RNA species. Most kits involve isolation of small RNAs by size fractionation using gel electrophoresis. Size fractionation of small RNAs requires involves running the total RNA on a gel, cutting a gel slice in the 14–30 nucleotide region, and purifying the gel slice. For higher concentrations of small RNAs, the excised gel slice can be concentrated by ethanol precipitation. An alternative to gel electrophoresis is the use of silica spin columns, which bind and elute small RNAs from a silica column. After isolation of small RNAs species from total RNA, the RNA is ready for cDNA synthesis and primer ligation.

cDNA Synthesis—

Universal to all RNA-Seq preparation methods is the conversion of RNA into cDNA because most sequencing technologies require DNA libraries. Most protocols for cDNA synthesis create libraries that were uniformly derived from each cDNA strand, thus representing the parent mRNA strand and its complement. In this conventional approach, the strand orientation of the original RNA is lost as the sequencing reads derived from each cDNA strand are indistinguishable in an effort to maximize efficiency of reverse transcription. However, strand information can be particularly valuable for distinguishing overlapping transcripts on opposite strands, which is critical for de novo transcript discovery (Parkhomchuk et al. 2009; Vivancos et al. 2010; Mills et al. 2013). Therefore, alternative library preparation protocols have since been developed that yield strand-specific reads. One strategy to preserve strand information is to ligate adapters in predetermined directions to single-stranded RNA or the first-strand of cDNA (Lister et al. 2008). Unfortunately, this approach is laborious and results in coverage bias at both the 5′ and 3′ ends of cDNA molecules. The preferred strategy to preserve strandedness is to incorporate a chemical label such as deoxy-UTP (dUTP) during synthesis of the second-strand cDNA that can be specifically removed by enzymatic digestion (Parkhomchuk et al. 2009). During library construction, this facilitates distinguishing the second-strand cDNA from the first strand. Although this approach is favored, the validity of antisense transcripts near highly expressed genes should be measured with caution because a small amount of reads (∼1%) have been observed from the opposite strand (Zeng and Mortazavi 2012).

Multiplexing—

Another consideration for constructing cost-effective RNA-Seq libraries is assaying multiple indexed samples in a single sequencing lane. The large number of reads that can be generated per sequencing run (e.g., a single lane of an Illumina HiSeq 2500 generates up to 750 million paired-end reads) permits the analysis of increasingly complex samples. However, increasingly high sequencing depths provide diminishing returns for lower complexity samples, resulting in oversampling with minimal improvement in data quality (Smith et al. 2010). Therefore, an affordable and efficient solution is to introduce unique 6-bp indices, also known as “barcodes,” to each RNA-Seq library. This enables the pooling and sequencing of multiple samples in the

same sequencing reaction because the barcodes identify which sample the read originated from. Depending on the application, adequate transcriptome coverage can be attained for 2–20 samples (Birney et al. 2007; Blencowe et al. 2009). To detect transcripts of moderate to high abundance, ~30–40 million reads are required to accurately quantify gene expression. To obtain coverage over the fullsequence diversity of complex transcript libraries, including rare and lowly-expressed transcripts, up to 500 million reads is required (Fu et al. 2014). As such, for any given study it is important to consider the level of sequencing depth required to answer experimental questions with confidence while efficiently using NGS resources.

Quantitative Standards

Although RNA-Seq is a widely used technique for transcriptome profiling, the rapid development of sequencing technologies and methods raises questions about the performance of different platforms and protocols. Variation in RNA-Seq data can be attributed to an assortment of factors, ranging from the NGS platform used to the quality of input RNA to the individual performing the experiment. To control for these sources of technical variability, many laboratories use positive controls or "spike-ins" for sequencing libraries. The External RNA Controls Consortium (ERCC) developed a set of universal RNA synthetic spike-in standards for microarray and RNA-Seq experiments (Jiang et al. 2011; Zook et al. 2012). The spike-ins consist of a set of 96 DNA plasmids with 273–2022 bp standard sequences inserted into a vector of ~2800 bp. The spike-in standard sequences are added to sequencing libraries at different concentrations to assess coverage, quantification, and sensitivity. These RNA standards serve as an effective quality control tool for separating technical variability from biological variability detected in differential transcriptome profiling studies.

Selection of Tissue or Cell Populations

When beginning an RNA-Seq experiment, one of the initial considerations is the choice of biological material to be used for library construction and sequencing. This choice is not trivial considering there are hundreds of cell types in over 200 different tissues that make up greater than 50 unique organs in humans alone. In addition to spatial (e.g., cell- and tissuetype) specificity, gene expression shows temporal specificity, such that different developmental stages will show unique expression signatures. Ultimately, the biological material chosen will be dependent on both the experimental goals and feasibility. For example, the tissue of choice for an investigation of unique gene expression signatures in colon cancer, the tissue choice is clear. However, for research studies investigating variation in gene expression across individuals in a population, the choice of biological material is less apparent and will likely depend on the feasibility of obtaining the biological samples (e.g., blood draws are less invasive and easier to perform than tissue biopsies).

Handling Tissue Heterogeneity—

Another consideration when selecting the biological source of RNA is the heterogeneity of tissues. The accuracy of gene expression quantification is dependent on the purity of samples. In fact, the heterogeneity can substantially impact estimations of transcript abundances in samples composed of multiple cell types. Most tissue samples isolated from the human body are heterogeneous by nature. Furthermore, pathological tissue samples are often composed of disease-state cells surrounded by normal cells. To isolate distinct cell types, experimental methods have been developed, including laser-capture microdissection and cell purification. Laser-capture microdissection enables the isolation of cell types that are morphologically distinguishable under direct microscopic visualization (Emmert-Buck et al. 1996). Although this technique yields high-quality RNA, the total yield is low and requires PCR amplification, thereby introducing amplification biases and creating less distinguishable expression profiles across different cell types (Kube et al. 2007). Cell purification and enrichment protocols are also available, such as differential centrifugation and fluorescence-activated cell sorting (Cantor et al. 1975). In conjunction with RNA-Seq, these experimental methods have overcome previous technical limitations and enable researchers to uncover unique expression signatures across specific cell-types and developmental stages (Moran et al. 2012; Nica et al. 2013). In addition to these experimental methods, in silico probabilistic models can be applied in downstream analysis to differentiate the transcript abundances of distinct cells from RNA-Seq data of heterogeneous tissue samples (Erkkila et al. 2010; Li and Xie 2013). Interestingly, in some cases, the sample heterogeneity can have advantages in transcriptome profiling by identifying novel pathways, implicating cellular origins of disease, or identifying previously unknown pathological sites (Alizadeh et al. 2000; Khan et al. 2001; Sorlie et al. 2001).

Single-Cell Transcriptomics—

Beyond tissue heterogeneity, considerable evidence indicates that cell-to-cell variability in gene expression is ubiquitous, even within phenotypically homogeneous cell populations (Huang 2009). Unfortunately, conventional RNA-Seq studies do not capture the transcriptomic composition of individual cells. The transcriptome of a single cell is highly dynamic, reflecting its functionality and responses to ever-changing stimuli. In addition to cellular heterogeneity resulting from regulation, individual cells show transcriptional “noise” that arises from the kinetics of mRNA synthesis and decay (Yang et al. 2003; Sun et al. 2012). Furthermore, genes that show mutually exclusive expression in individual cells may be observed as genes showing co-expression in expression analyses of bulk cell populations. To uncover cell-to-cell variation within populations, significant efforts have been invested in developing single-cell RNA-Seq methods. The biggest challenge has been extending the limits of library preparation to accommodate extremely low input RNA. A human cell contains RNA-to-cDNA conversion is imperfect, estimated to be as low as 5%–25% of all transcripts (Islam et al. 2012). In addition, PCR amplification methods do not linearly amplify transcript and are prone to introduce biases based on the nucleic acid composition of different transcripts, ultimately altering the relative abundance of these transcripts in the sequencing library. Methods that avoid PCR amplification

steps, such as CEL-Seq, through linear in vitro amplification of the transcriptome can avoid these biases (Hashimshony et al. 2012). In addition, the use of nanoliter-scale reaction volumes with microfluidic devices as opposed to microliter-scale reactions can reduce biases that arise during sample preparation (Wu et al. 2014). Although single-cell methods are still under active development, quantitative assessments of these techniques indicate that obtaining accurate transcriptome measurements by single-cell RNA-Seq is possible after accounting for technical noise (Brennecke et al. 2013; Wu et al. 2014). These methods will undoubtedly be important for uncovering oscillatory and heterogeneous gene expression within single-cell types, as well as identifying cell-specific biomarkers that further our understanding of biology across many physiological and pathological conditions.

Sequencing Platforms for Transcriptomics

When designing an RNA-Seq experiment, the selection of a sequencing platform is important and dependent on the experimental goals. Currently, several NGS platforms are commercially available and other platforms are under active technological development (Metzker 2010). The majority of high-throughput sequencing platforms use a sequencingby-synthesis method to sequence tens of millions of sequence clusters in parallel. The NGS platforms can often be categorized as either ensemble-based (i.e. sequencing many identical copies of a DNA molecule) or single-molecule-based (i.e. sequencing a single DNA molecule). The differences between these sequencing techniques and platforms can affect downstream analysis and interpretation of the sequencing data. In recent years, the sequencing industry has been dominated by Illumina, which applies an ensemble-based sequencing-by-synthesis approach (Bentley et al. 2008). Using fluorescently labeled reversible-terminator nucleotides, DNA molecules are clonally amplified while immobilized on the surface of a glass flowcell. Because molecules are clonally amplified, this approach provides the relative RNA expression levels of genes. To remove potential PCR-amplification biases, PCR controls and specific steps in the downstream computational analysis are required. One major benefit of ensemble-based platforms is low sequencing error rates (<1%) dominated by single mismatches. Low error rates are particularly important for sequencing miRNAs, whose relatively small sizes result in misalignment or loss of reads if error rates are too high. Currently, the Illumina HiSeq platform is the most commonly applied next-generation sequencing technology for RNA-Seq and has set the standard for NGS sequencing. The platform has two flow cells, each providing eight separate lanes for sequencing reactions to occur. The sequencing reactions can take between 1.5 and 12 d to complete, depending on the total read length of the library. Even more recently, Illumina released the MiSeq, a desktop sequencer with lower throughput but faster turnaround (generates ~30 million paired-end reads in 24 h). The simplified workflow of the MiSeq instrument offers rapid turnaround time for transcriptome sequencing on a smaller scale.

Single-molecule-based platforms such as PacBio enable single-molecule real-time (SMRT) sequencing (Eid et al. 2009). This approach uses DNA polymerase to perform uninterrupted template-directed synthesis using fluorescently labeled nucleosides. As each base is enzymatically incorporated into a growing DNA strand, a distinctive pulse of fluorescence is detected in real-time by zero-mode waveguide nanostructure arrays. An advantage of SMRT is that it does not include a PCR amplification step, thereby avoiding amplification bias and improving uniform coverage across the transcriptome. Another advantage of this sequencing approach is the ability to produce extraordinarily long reads with average lengths of 4200 to 8500 bp, which greatly improves the detection of novel transcript structures (Au et al. 2013; Sharon et al. 2013). A critical disadvantage of SMRT is a high rate of errors (∼5%) that are predominately characterized by insertions and deletions (Carneiro et al. 2012); the high error rate results in misalignment and loss of sequencing reads due to the difficulty of matching erroneous reads to the reference genome.

Another important consideration for choosing a sequencing platform is transcriptome assembly. Transcriptome assembly, which is discussed in greater detail later, is necessary to transform a collection of short sequencing reads into a set of full-length transcripts. In general, longer sequencing reads make it simpler to accurately and unambiguously assemble transcripts, as well as identify splicing isoforms. The extremely long reads generated by the PacBio platform are ideal for de novo transcriptome assembly in which the reads are not aligned to a reference transcriptome. The longer reads will facilitate an accurate detection of alternative splice isoforms, which may not be discovered with shorter reads. Moleculo, a company acquired by Illumina, has developed long-read sequencing technology capable of producing 8500 bp reads. Although it has yet to be widely adopted for transcriptome sequencing, the long reads aid transcriptome assembly. Lastly, Illumina has developed protocols for its desktops MiSeq to sequence slightly longer reads (up to 350 bp). Although much shorter than PacBio and Moleculo reads, the longer MiSeq reads can also be used to improve both de novo and reference transcriptome assembly

Transcriptome Analysis

Gene expression profiling by RNA-Seq provides an unprecedented high-resolution view of the global transcriptional landscape. As the sequencing technologies and protocol methodologies continually evolve, new informatics challenges and applications develop. Beyond surveying gene expression levels, RNA-Seq can also be applied to discover novel gene structures, alternatively spliced isoforms, and allele-specific expression (ASE). In addition, genetic studies of gene expression using RNA-Seq have observed genetically correlated variability in expression, splicing, and ASE (Montgomery et al. 2010; Pickrell et al. 2010; Battle et al. 2013; Lappalainen et al. 2013). This section will introduce how expression data are analyzed to provide greater insight into the extensive complexity of transcriptomes.

RNA-Sequencing Data Analysis Workflow

The conventional pipeline for RNA-Seq data includes generating FASTQ-format files contains reads sequenced from an NGS platform, aligning these reads to an annotated reference genome, and quantifying expression of genes (Fig. 2). Although basic sequencing analysis tools are more accessible than ever, RNA-Seq analysis presents unique computational challenges not encountered in other sequencing-based analyses and requires specific consideration to the biases inherent in expression data.

Read Alignment—Mapping RNA-Seq reads to the genome is considerably more challenging than mapping DNA sequencing reads because many reads map across splice junctions. In fact, conventional read mapping algorithms, such as Bowtie (Langmead et al. 2009) and BWA (Li and Durbin 2009), are not recommended for mapping RNA-Seq reads to the reference genome because of their inability to handle spliced transcripts. One approach to resolving this problem is to supplement the reference genome with sequences derived from exon–exon splice junctions acquired from known gene annotations (Mortazavi et al. 2008). A preferred strategy is to map reads with a "splicing-aware" aligner that can recognize the difference between a read aligning across an exon–intron boundary and a read with a short insertion. As RNA-Seq data have become more widely used, a number of splicing-aware mapping tools have been developed specifically for mapping transcriptome data. The more commonly used RNA-Seq alignment tools include GSNAP (Wu and Nacu 2010), MapSplice (Wang et al. 2010a), RUM (Grant et al. 2011), STAR (Dobin et al. 2013), and TopHat (Trapnell et al. 2009) (Table 2). Each aligner has different advantages in terms of performance, speed, and memory utilization. Selecting the best aligner to use depends on these metrics and the overall objectives of the RNA-Seq study. Efforts to systematically evaluate the performance of RNA-Seq aligners have been initiated by GENCODE's RNASeq Genome Annotation Assessment Project3 (RGASP3), which has found major performance difference between alignments tools on numerous benchmarks, including alignment yield, basewise accuracy, mismatch and gap placement, and exon junction discovery (Engstrom et al. 2013).

Transcript Assembly and Quantification—

After RNA-Seq reads are aligned, the mapped reads can be assembled into transcripts. The majority of computational programs infer transcript models from the accumulation of read alignments to the reference genome (Trapnell et al. 2010; Li et al. 2011; Roberts et al. 2011a; Mezlini et al. 2013) (Table 2). An alternative approach for transcript assembly is de novo reconstruction, in which contiguous transcript sequences are assembled with the use of a reference genome or annotations (Robertson et al. 2010; Grabherr et al. 2011; Schulz et al. 2012). The reconstruction of transcripts from short-read data is a major challenge and a gold standard method for transcript assembly does not exist. The nature of the transcriptome (e.g., gene complexity, degree of polymorphisms, alternative splicing, dynamic range of expression), common technological challenges (e.g., sequencing errors), and features of the bioinformatics workflow (e.g., gene annotation, inference of isoforms) can substantially affect transcriptome assembly quality. RGASP3 has initiated efforts to evaluate computational methods for

transcriptome reconstruction and has found that most algorithms can identify discrete transcript components, but the assembly of complete transcript structures remains a major challenge (Steijger et al. 2013).

A common downstream feature of transcript reconstruction software is the estimation of gene expression levels. Computational tools such as Cufflinks (Trapnell et al. 2010), FluxCapacitor (Montgomery et al. 2010; Griebel et al. 2012), and MISO (Katz et al. 2010), quantify expression by counting the number of reads that map to full-length transcripts (Table 2). Alternative approaches, such as HTSeq, can quantify expression without assembling transcripts by counting the number of reads that map to an exon (Anders et al. 2013). To accurately estimate gene expression, read counts must be normalized to correct for systematic variability, such as library fragment size, sequence composition bias, and read depth (Oshlack and Wakefield 2009; Roberts et al. 2011b). To account for these sources of variability, the reads per kilobase of transcripts per million mapped reads (RPKM) metric normalizes a transcript's read count by both the gene length and the total number of mapped reads in the sample. For paired end-reads, a metric that normalizes for sources of variances in transcript quantification is the paired fragments per kilobase of transcript per million mapped reads (FPKM) metric, which accounts for the dependency between paired-end reads in the RPKM estimate (Trapnell et al. 2010). Another technical challenge for transcript quantification is the mapping of reads to multiple transcripts that are a result of genes with multiple isoforms or close paralogs. One solution to correct for this "read assignment uncertainty" is to exclude all reads that do not map uniquely, as in Alexa-Seq (Griffith et al. 2010). However, this strategy is far from ideal for genes lacking unique exons. An alternative strategy used by Cufflinks (Trapnell et al. 2012), and MISO (Katz et al. 2010) is to construct a likelihood function that models the sequencing experiment and estimates the maximum likelihood that a read maps to a particular isoform.

Considerations for miRNA Sequencing Analysis—

The general approach for analysis of miRNA sequencing data is similar to approaches discussed for mRNA. To identify known miRNAs, the sequencing reads can be mapped to a specific database, such as miRBase, a repository containing over 24,500 miRNA loci from 206 species in its latest release (v21) in June 2014 (Kozomara and Griffiths-Jones 2014). In addition, several tools have been developed to facilitate analysis of miRNAs including the commonly used tools miRanalyzer (Hackenberg et al. 2011) and miRDeep (An et al. 2013). MiRanalyzer can detect known miRNAs annotated on miRBase as well as predict novel miRNAs using a machine-learning approach based on the random forest method with a broad range of features. Similarly, miRDeep is able to identify known miRNAs and predict novel miRNAs using properties of miRNA biogenesis to score the compatibility of the position and frequency of sequenced RNA from the secondary structure of precursor miRNAs. Although miRDeep and miRanalyzer contain modules for target prediction, expression quantification, and differential expression, the methods developed for mRNA quantification and differential expression can also be applied to miRNA data (Eminaga et al. 2013).

Quality Assessment and Technical Considerations

At each stage in the RNA-Seq analysis pipeline, careful consideration should be applied to identifying and correcting for various sources of bias. Bias can arise throughout the RNA-Seq experimental pipeline, including during RNA extraction, sample preparation, library construction, sequencing, and read mapping (Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012; 't Hoen et al. 2013). First, the quality of the raw sequence data in FASTQ-format files should be evaluated to ensure high-quality reads. User-friendly software tools designed to generate quality overviews include the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit), the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), and the RobiNA package (Lohse et al. 2012). Several important parameters that should be evaluated include the sequence diversity of reads, adaptor contamination, base qualities, nucleotide composition, and percentage of called bases. These technical artifacts can arise at the sequencing stage or during the construction of the RNA-Seq. For example, the 5′ read end, derived from either end of a double-stranded cDNA fragment, shows higher error rate due to mispriming events introduced by the random oligos during the RNA-Seq library construction protocol (Lin et al. 2012). If possible, actions to correct for these biases should be performed, such as trimming the ends of reads, to expedite the speed and improve the quality of the read alignments. After aligning the reads, additional parameters should be assessed to account for biases that arise at the read mapping stage. These parameters include the percentage of reads mapped to the transcriptome, the percentage of reads with a mapped mate pair, the coverage bias at the 5′- and 3′-ends, and the chromosomal distribution of reads.

One of the most common sources of mapping errors for RNA-Seq data occurs when a read spans the splicing junction of an alternatively spliced gene. A misalignment can be easily introduced due to ambiguous mapping of the read end to one of the two (or more) possible exons and is especially common when reads are mapped to a reference transcriptome that contains an incomplete annotation of isoforms (Kleinman and Majewski 2012; Pickrell et al. 2012). If genotype information is available, the integrity of the samples should also be evaluated by investigating the correlation of single-nucleotide variants (SNVs) between the DNA and RNA reads ('t Hoen et al. 2013). The concordance between the DNA and RNA sequencing data may provide insight into sample swaps or sample mixtures caused accidentally as a result of personnel or equipment error. In the case of a swapped sample, more discordant variants would be observed between the DNA and RNA sequencing data. In the case of a mixture of samples, more significant patterns of allele-specific expression would be observed than expected for a single individual as a result of more combinations of heterozygous and homozygous sites that would skew the alleles beyond the expected 1:1 allelic ratio.

Differential Gene Expression

A primary objective of many gene expression experiments is to detect transcripts showing differential expression across various conditions. Extensive statistical approaches have been developed to test for differential expression with microarray data, where the continuous probe intensities across replicates can be approximated by a normal distribution (Cui and Churchill 2003; Smyth 2004; Grant et al. 2005). Although in principle these approaches are also applicable to RNA-Seq data, different statistical models must be considered for discrete read counts that do not fit a normal distribution. Early RNA-Seq studies suggested that the distribution of read counts across replicates fit a Poisson distribution, which formed the basis for modeling RNA-Seq count data (Marioni et al. 2008). However, further studies indicated that biological variability is not captured by the Poisson assumption, resulting in high falsepositive rates due to underestimation of sampling error. Hence, negative binomial distribution models that take into account overdispersion or extra-Poisson variation have been shown to best fit the distribution of read counts across biological replicates.

To model the count-based nature of RNA-Seq data, complex statistical models have been developed to handle sources of variability that model overdispersion across technical and biological replicates. One source of variability is differences in sequencing read depth, which can artificially create differences between samples. For instance, differences in read depth will result in the samples appearing more divergent if raw read counts between genes are compared. To correct for this, it is advantageous to transform raw read count data to FPKM or RPKM values in differential expression analyses. Although this correction metric is commonly used in place of read counts, the presence of several highly expressed genes in a particular sample can significantly alter the RPKM and FPKM values. For example, a highly expressed gene can "absorb" many reads, consequently repressing the read counts for other genes and artificially inflating gene expression variation. To account for this bias, several statistical models have been proposed that use the highly expressed genes as model covariates (Robinson and Oshlack 2010). Another source of variability that has been observed is that the distribution of sequencing reads is unequal across genes. Therefore, a two-parameter generalized Poisson model that simultaneously considers read depth and sequencing bias as independent parameters was developed and shown to improve RNA-Seq analysis (Srivastava and Chen 2010).

More complex normalization methods have also been developed to account for hidden covariates without removing significant biological variability. For example, the probabilistic estimation of expression residuals (PEER) framework (Stegle et al. 2012) and the hidden covariates with prior (HCP) framework (Mostafavi et al. 2013) are methods that use a Bayesian approach to infer hidden covariates and remove their effects from expression data. To detect differential expression, a variety of statistical methods have been designed specifically for RNA-Seq data. A popular tool to detect differential expression is Cuffdiff, which is part of the Tuxedo suite of tools (Bowtie, Tophat, and Cufflinks) developed to analyze RNA-Seq data (Trapnell et al. 2013). In addition to Cuffdiff, several other packages support testing differential expression, including baySeq (Hardcastle and Kelly 2010), DESeq (Anders and Huber 2010), DEGseq

(Wang et al. 2010b), and edgeR (Robinson et al. 2010) (Table 2). Although these packages can assign significance to differentially expressed transcripts, the biological observations should be carefully interpreted. Each model makes specific assumptions that may be violated in the context of the observed data; therefore, an understanding of the model parameters and their constraints is critical for drawing meaningful and accurate biological conclusions (Bullard et al. 2010). Furthermore, replicates in RNA-Seq experiments are crucial for measuring variability and improving estimations for the model parameters (Tarazona et al. 2011; Glaus et al. 2012). Biological replicates (e.g., cells grown on two different plates under the same conditions) are preferred to technical replicates (e.g., one RNA-Seq library sequenced on two different lanes), which show little variation. Although the number of replicates required per condition is an open research question, a minimum of three replicates per sample has been suggested (Auer and Doerge 2010). In many cases, multiplexed RNA-Seq libraries can be used to add biological replicates without increasing sequencing costs (if sequenced at a lower depth) and will greatly improve the robustness of the experimental design (Liu et al. 2014). Additionally, the accuracy of measurements of differential gene expression can be further improved by using ERCC spike-in controls to distinguish technical variation from biological variation.

Allele-Specific Expression

A major advantage of RNA-Seq is the ability to profile transcriptome dynamics at a singlenucleotide resolution. Therefore, the sequenced transcript reads can provide coverage across heterozygous sites, representing transcription from both the maternal and paternal alleles. If a sufficient number of reads cover a heterozygous site within a gene, the null hypothesis is that the ratio of maternal to paternal alleles is balanced. Significant deviation from this expectation suggests allele-specific expression (ASE). Potential mechanisms for ASE include genetic variation (e.g., single-nucleotide polymorphism in a cis-regulatory region upstream of a gene) and epigenetic effects (e.g., genomic imprinting, methylation, histone modifications, etc.). Early studies showed that allele-specific differences can affect up to 30% of loci within an individual (Ge et al. 2009) and are caused by both common and rare genetic variants (Pastinen 2010). Studies have also applied ASE to identify expression modifiers of protein-coding variation (Lappalainen et al. 2011; Montgomery et al. 2011), effects of loss-of-function variation (MacArthur et al. 2012), and differences between pathogenic and healthy tissues (Tuch et al. 2010). Furthermore, ASE studies using singlecell transcriptomics have uncovered a stochastic pattern of allelic expression that may contribute to variable expressivity, a novel perspective which may have fundamental implications for variable disease penetrance and severity (Deng et al. 2014).

Conventional workflows to detect ASE involve counting reads containing each allele at heterozygous sites and applying a statistical test, such as the binomial test or the Fisher's exact test (Degner et al. 2009; Rozowsky et al. 2011; Wei and Wang 2013). However, more rigorous statistical approaches are necessary to overcome technical challenges involved in ASE detection. These challenges include read-mapping bias, sampling variance, overdispersion at extreme read

depths, alternatively spliced alleles, insertions and deletions (indels), and genotyping errors. To account for overdispersion, one approach is to model allelic read counts using a beta-binomial distribution at individual loci (Sun 2012); however, accurate estimation of the overdispersion parameter requires replicates and, in our experience, major source of bias come from site-specific mapping differences. Another strategy is to use a hierarchical Bayesian model that combines information across loci, as well as across replicates and technologies, to make global and site-specific inferences for ASE (Skelly et al. 2011). To assess reference-allele mapping bias, the number of mismatches in reads containing the nonreference allele should be assessed as increased bias is observed with greater sequence divergence between alleles (Stevenson et al. 2013). To correct for read-mapping bias, an enhanced reference genome can be constructed that masks all SNP positions or includes the alternative alleles at polymorphic loci (Degner et al. 2009; Satya et al. 2012). Statistical methods to better address these technical biases are under active development and are expected to foster further improvements in ASE detection

Expression Quantitative Trait Loci

Another prominent direction of RNA-Seq studies has been the integration of expression data with other types of biological information, such as genotyping data. The combination of RNA-Seq with genetic variation data has enabled the identification of genetic loci correlated with gene expression variation, also known as expression quantitative trait loci (eQTLs). This expression variation caused by common and rare variants is postulated to contribute to phenotypic variation and susceptibility to complex disease across individuals (Majewski and Pastinen 2011). The goal of eQTL analysis is to identify associations that will uncover underlying biological processes, discover genetic variants causing disease, and determine causal pathways. Initial eQTL studies using RNA-Seq data identified a greater number of statistically significant eQTLs than had been identified by microarray studies (Montgomery et al. 2010; Pickrell et al. 2010). Most of the eQTLs identified directly influenced gene expression in an allele-specific manner and were located near transcriptional start sites, indicating that eQTLs could modulate expression directly, or in cis. Later studies identified trans-eQTLs, which are variants that affect the expression of a distant gene (>1 Mb) by modifying the activity or expression of upstream factors that regulate the gene (Fehrmann et al. 2011; Battle et al. 2013; Westra et al. 2013). Although trans-eQTLs show weaker effects and present validation difficulties, they can potentially reveal previously unknown pathways in gene regulation networks.

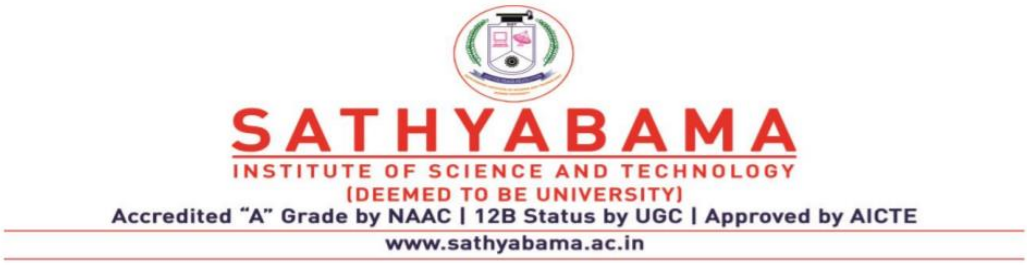
RNA-Seq has revolutionized QTL analyses because it enables association analyses of more than just gene expression levels alone. For example, RNA-Seq provides unprecedented opportunity to investigate variations in splicing by profiling alternately spliced isoforms of a gene. This has enabled the identification of variants influencing the quantitative expression of alternatively spliced isoforms commonly referred to as splicing-QTLs (sQTLs) (Lalonde et al. 2011). In addition, specific RNA-Seq library constructions (e.g., ribo-depleted) have enabled the detection of eQTLs affecting other RNA species; recent studies have identified variants affecting the expression of various ncRNAs, including long intergenic noncoding RNAs (Montgomery et al.

2010; Gamazon et al. 2012; Kumar et al. 2013; Popadin et al. 2013). The expanding potential of RNA-Seq to associate phenotypic variations with genetic variation offers an enhanced understanding of gene regulation.

Traditional eQTL mapping methods that were developed for microarray data use linear models such as linear regression and ANOVA to associate genetic variants with gene expression (Kendziorski and Wang 2006). These methods have been directly applied to RNA-Seq data following appropriate normalization of total read counts. Most eQTL studies perform separate testing for each transcript-SNP pair using linear regression and ANOVA models to detect significant association. Nonlinear approaches have also been developed to test associations, such as generalized linear and mixed models, Bayesian regression (Servin and Stephens 2007). Alternative models, such as Merlin, have also been developed to detect eQTLs from expression data that include related individuals using pedigree data (Abecasis et al. 2002). In addition, several methods have been developed to simultaneously test the effect of multiple SNPs on the expression of a single gene using Bayesian methods (Lee et al. 2008). To further improve on the detection of causal regulatory variants, several studies have integrated ASE information with eQTL analysis. These studies showed that genetic variants showing allele-specific effects and identified as eQTLs show higher enrichment in functional annotations and provide stronger evidence of cis-regulatory impact (Battle et al. 2013; Lappalainen et al. 2013; Sun and Hu 2013). Because high-throughput sequencing has created genotype data sets featuring millions of SNPs and expression data sets featuring tens of thousands of transcripts, the task of testing billions of transcript-SNP pairs in eQTL analysis can be computationally intensive. To mitigate this computational burden, software has been developed such as Matrix eQTL to efficiently test the associations by modeling the effect of genotype as either additive linear (least squares model) or categorical (ANOVA model) (Shabalin 2012). Because of the large number of tests performed, it is important to correct for multiple-testing by calculating the false discovery rate (Benjamini and Hochberg 1995; Yekutieli and Benjamini 1999) or resampling using bootstrap or permutation procedures (Karlsson 2006; Zhang et al. 2012).

However, the design and interpretation of eQTL studies is not straightforward. Many complications result from the complexity of gene regulation, which shows both spatial (cell and tissue location) specificity as well as temporal (developmental stage) specificity. For instance, several studies have performed eQTL analysis across multiple tissues, indicating that genetic regulatory elements can have tissue-specific effects (Petretto et al. 2006; Schadt et al. 2008; Dimas et al. 2009; Kwan et al. 2009; Grundberg et al. 2012; Flutre et al. 2013). Therefore, future eQTL analyses should test for SNP-transcript associations in well-defined cell types that are relevant to the trait of interest (Lonsdale et al. 2013). For example, a study detecting eQTLs in cardiovascular disease should use heart tissue while a study interested in autoimmune disease should use whole blood. Another major consideration for eQTL studies is accounting for population structure and elucidating the causal variants (Stranger et al. 2012). The structure of genomic variation can vary significantly between populations and will influence the resolution of

any genetic association study (Frazer et al. 2007; Altshuler et al. 2010). Furthermore, if substantial linkage disequilibrium (LD) exists within the genome, the associated genetic variant is often “tagging” the causal variant rather than acting as the causal regulatory variant itself. As eQTL studies integrate data across different populations and use population-scale genome sequencing, the ability to elucidate causal variants will greatly improve



SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – IV - Next Generation Sequencing– SBI1606

Chromatin immunoprecipitation followed by sequencing (ChIPseq) analysis is a key technology in epigenomic research. This method uses an antibody for a specific DNA-binding protein or a histone modification to identify enriched loci within a genome [1,2]. Histone modifications are used in the ChIP-seq analysis field to dissect the characteristics and the biological functions of epigenetic signatures. Advances in next-generation sequencing (NGS) technology and computational analysis enable us to systematically understand how the epigenomic landscape contributes to cell identity [3], development [4], lineage specification [5–8], cancer [9], and other diseases [10,11].

Five "core histone marks", proposed by Roadmap Epigenomics Consortium [12], are widely used for ChIP-seq analysis:

- H3 lysine 4 monomethylation (H3K4me1) or H3 lysine 27 acetylation (H3K27ac), which is associated with enhancer regions;
- H3 lysine 4 trimethylation (H3K4me3), which is associated with promoter regions;
- H3 lysine 36 trimethylation (H3K36me3), which is associated with transcribed regions in gene bodies;
- H3 lysine 27 trimethylation (H3K27me3), which is associated with Polycomb repression; and
- H3 lysine 9 trimethylation (H3K9me3), which is associated with heterochromatin.

In addition to genome-wide identification of specific epigenome marks (e.g., enhancers) in a specific cell-line, core histone mark enrichment profiles are used to segment and annotate whole-genome regions into distinct "chromatin states," which represent more detailed characteristic epigenetic signatures (e.g., weak transcription and poised promoter). Maturation of high-quality ChIP-seq databases by large consortia such as ENCODE, the Roadmap Epigenomics Consortium, and the International Human Epigenome Consortium (IHEC) accelerate chromatin state annotation for various cell lines and tissues. Many studies leverage the accumulated epigenomic information to infer additional genome dynamics using machine-learning approaches.

In this review, we first address the major steps in a typical ChIP-seq computational analysis workflow. Because there are numerous important studies in this field, we focus on outlining the concept for each step by referencing previous important reviews instead of describing each method. Next, we introduce several advanced ChIP-seq applications for histone modifications, including prediction of gene expression level and enhancer-promoter looping, and data imputation. Finally, we discuss recently developed methodologies for single-cell ChIP-seq (scChIP-seq) analysis that elucidate the cellular diversity within complex tissues and cancers.

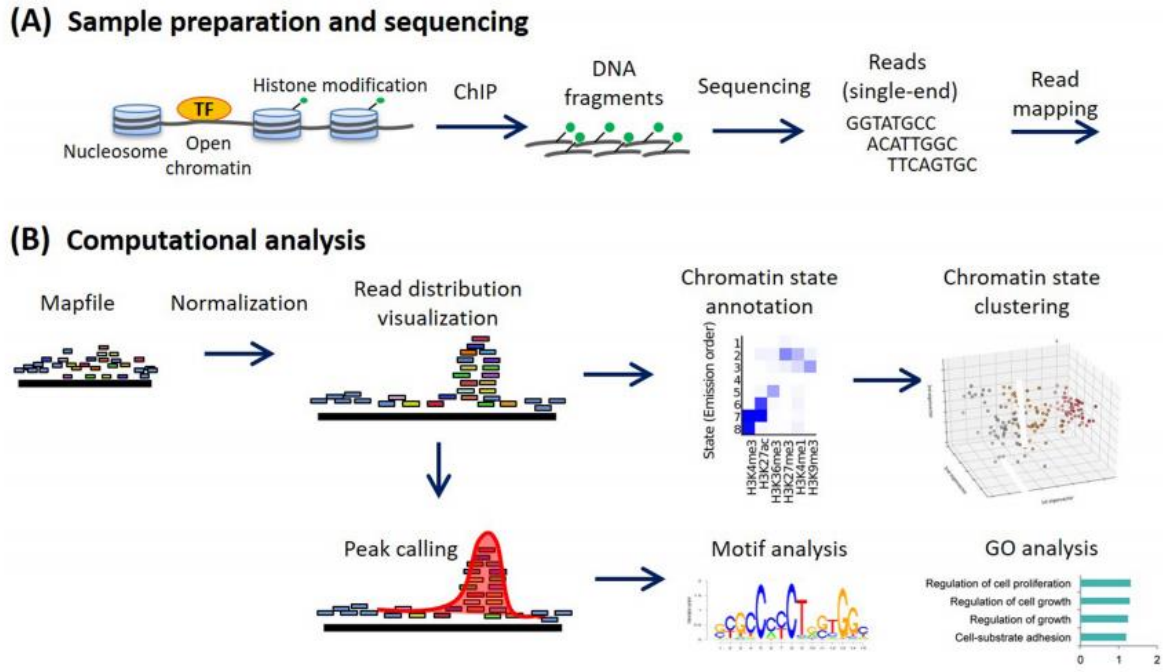


Fig. 1. ChIP-seq analysis workflow. (A) Sample preparation and sequencing. (B) Computational analysis in a canonical ChIP-seq analysis. Various analyses are implemented using normalized read distribution.

2. ChIP-seq analysis workflow

In this section, we describe the step-by-step workflow of a typical ChIP-seq analysis (Fig. 1). Also, see our previous review [16] for details and considerations for each step.

2.1. Environmental setup

Computational tools for NGS analysis are written in various computational languages such as C++, R, Python, Java, and Perl. Each language requires a different setup method. While most are executed on Linux systems, Mac terminal and Windows Subsystem for Linux (WSL) can also be used (note that some specific errors might occur due to different library names and dependencies). One major upcoming issue is that Python2 will not be maintained after 2020 (https://www.python.org/doc/sunset-python-2/). There are several tools that require Python2 but have not been updated to Python3 (e.g., Peakzilla [17] and ChromTIME [18]). In the near future, users will have to consider replacing these packages for newer alternatives. If users want to keep using these older applications (e.g., because current analysis pipelines for big projects are difficult to modify), virtual environments like Docker (https://www.docker.com/) or Singularity (https://sylabs.io/) provide a secure, isolated analysis environment. Several computational tools and analysis pipelines are released as Docker images, which are downloadable pre-compiled computational environments. These images remove difficult extraction and installation.

2.2. Downloading ChIP-seq data from public databases

Multiple public databases are available to download ChIP-seq data of histone modifications (Table 1). We recently published an epigenome database for human endothelial cells (entry 4 in Table 1), which contains 424 histone modification ChIP-seq and 67 RNA sequencing (RNA-seq) datasets obtained from nine blood vessel types from the human body [19]. Various data types are available (e.g., reads, mapfiles, bigwig files, and peak lists), that are suitable as ChIP-seq analysis tutorial data.

2.3. Technical considerations of ChIP-seq analysis for histone modifications

The reliability of a ChIP analysis is governed by antibody quality, including specificity and signal-to-noise ratio (S/N) [20]. Since the false-positive enriched sites derived from nonspecific antibody-DNA binding may confound the analysis, unexpected ChIP-seq results should be validated using multiple antibodies [21].

While most ChIP-seq tools are designed for sharp peaks that are located at specific genomic positions, such as transcription start sites (TSSs), some histone modifications are associated with large genomic domains, resulting in broadly distributed enrichment regions [1]. H3K27me3 and H3K36me3 enrichments distribute across several hundred kilobases, while H3K9me3 peaks often expand to a few megabases. The enhancer markers, H3K27ac and H3K4me1, produce sharp peaks, but sometimes construct broadly enriched regions called "super enhancers" [22]. H3K4me3 promoter markers can also cover broad domains in mouse oocytes [23]. This peak shape and broadness variation affects the choice of optimal computational tools. For example, ROSE [24] is specifically designed to detect super-enhancer sites, which

Table 1
Public ChIP-seq databases.

Database	URL	Reference
ENCODE portal	https://www.encodeproject.org/	[118]
ROADMAP epigenome database	http://www.roadmapepigenomics.org/	[12]
IHEC Data Portal	https://epigenomesportal.ca/ihec/	[119]
Epigenome database for human endothelial cells	https://rnakato.github.io/HumanEndothelialEpigenome/	[19]

congregate multiple enhancer sites close together. Music [25] can estimate the average sample peak width to be investigated.

Read mapping

The sequenced reads (FASTQ or CSFSATQ format) are mapped using tools such as Bowtie [26], Bowtie2 [27], or BWA [28]. Bowtie2 and BWA can consider indels (insertions and deletions) by gapped alignments, which is appropriate for long and/or paired-end reads (see [29] for a comparison of mapping tools and parameters). There are several output formats for map files, such as SAM, BAM, CRAM and tagAlign. While the BAM format is the most widely used so far, the more spaceefficient CRAM format is maturing and will likely be the next standard (https://www.ga4gh.org/cram/). After alignment, reads mapped to the same genomic positions are filtered as redundant reads, and the remaining nonredundant reads are used for analysis.

Peak calling

The peak-calling step identifies significantly enriched loci (peaks) in the genome. Peak-calling results are generally returned in BED format. Although ChIP-seq peaks do not have strand information, it can be estimated from the gene information when focusing on the histone marks that are enriched around TSS, for instance. While MACS2 [30] is the most commonly used peak-calling tool, numerous peak-calling tools were recently developed (see [16,31,32] for reviews). However, no tool can achieve 100% accuracy. Therefore, a practical strategy is to obtain a large number of peaks with a relaxed threshold that contain true positives and noise, and then extract subgroups using another way to improve specificity, e.g. selecting consistent signal among biological replicates using the Irreproducible Discovery Rate (IDR)

ChIP-seq data quality assessment

Quality check (QC) of ChIP-seq samples is critical to judge whether sequencing data are of high quality and suitable for further analyses. Various quantitative QC measures have been developed [16,20]. Among them, the particularly important metrics are:

- Mapping ratio, which reflects read quality and the proportion of sequenced reads that are derived from true genomic DNA. For example, the mapping ratio for samples sequenced by Illumina HiSeq System (e.g., Hiseq2500) should be over 80%. The exception is a sample for non-DNA-binding proteins such as IgG, which often has a lower mapping ratio (~60%).
- Read depth (the number of nonredundant mapped reads). Sufficient read depth depends on the genome size and the antibody S/N ratio [1]. The ENCODE consortium suggested at least 10 million uniquely mapped reads as a minimum to analyze sharp-mode peaks of human samples [20]. Broad histone marks often have weaker S/N and require more reads (e.g., > 40 million for human) as a practical minimum for peak calling [33].

- Library complexity (the proportion of nonredundant reads). It ranges from 0 to 1.0, and the ENCODE consortium suggested the complexity > 0.8 for 10 million mapped reads [20]. Lower values (less than 0.6) indicate excessive PCR amplification from a small amount of initial DNA [16].
- The normalized strand coefficient (NSC, obtained by SSP [34]), a S/ N indicator for both sharp and broad marks (phantompeakqualtools [20] can only calculate NSC for sharp marks). In-depth validation using > 1,000 publicly available ChIP-seq datasets for multiple species suggested that the recommended threshold value is NSC > 5.0 and NSC > 1.5 for sharp and broad marks, respectively [34]. Input samples should have a low S/N and therefore NSC values should be < 2.0.
- Background uniformity (Bu) [34]. Bu reflects the read distribution bias in background regions and ranges from 0 to 1.0. Low values (less than 0.8) suggest that the read distribution is more congregated or biased than expected, resulting in numerous false positives in obtained peaks [35]. For the genome that has extensive copynumber variations (e.g., MCF-7 cells), a relaxed threshold value (> 0.6) is desirable.
- GC summit bias, reflecting biases during immunoprecipitation and PCR amplification [35]. In general, the GC summit of typical ChIPseq data becomes similar to the reference genome (e.g., ∼50% for human [19]). Unexpected GC-rich summit (e.g., over 60% for human) is often manifested due to PCR amplification biases [35] and/or false-positive peaks derived from 'hyper-ChIPable' regions associated with CpG islands

Visualization

Having developed various statistical methods and quality metrics for ChIP-seq data, visual inspection of read distribution is effective to intuitively assess and analyze the obtained data, e.g., detecting suspicious peaks derived from hyper-ChIPable regions [36]. For that, interactive visualization tools such as Integrated Genome Viewer (IGV) [38] or SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/ seqmonk/) are available. Several web servers (e.g. UCSC genome browser [39] and WashU Epigenome Browser [40]) can integrate the obtained ChIP-seq results with other annotation data, such as evolutionary conservation and gene expression in various tissues.

Normalization for comparative analysis

Read normalization is essential to mitigate technical variance before comparative analysis [35]. Simple total read normalization is commonly used, which scales the sample read number to be the same. The underlying assumption is that the difference in mapped reads among samples is sufficiently smaller than the total read number. This assumption is not always satisfied, and therefore, several methods have been developed to identify differentially enriched regions between two conditions, some of which are specifically designed for histone modification data [41,42]. Since the obtained results vary considerably among tools due to the underlying statistical assumptions, the choice of method will crucially

impact the outcome [43]. Quantitative comparison across more than two groups is more complicated. When the expected S/N value is similar among samples, statistical methods for differential gene expression analysis can be used [44]. It is also possible to utilize quantile normalization [19] when the S/N for most common peaks is similar among samples (e.g., a single antibody for all samples). If the S/N highly varies among samples (e.g., between with and without stimulation), consider spike-in analysis (also called calibration analysis) [45,46]. This method is a wet-based solution that adds the same amount of DNA from a different species to all samples before or after immunoprecipitation and estimates the weight coefficient based on the number of derived reads. In contrast to computational normalization methods that are limited to relative differences, spike-in ChIP-seq enables investigation of absolute-level differences [16]. However, quantitative ChIP-seq comparisons are still often confounded by intrinsic noisiness and variability caused by multi-step sample preparation, even after normalization [43]. In this case, simple binary comparisons (identifying common or unique peaks) might be desirable, though some false positiveS/Negatives will likely occur in the obtained results.

Functional analysis

Motif analysis investigates the sequence specificity inherent in called peaks or specific epigenome regions (e.g., enhancer sites), and estimates the likely transcription factor binding sites within identified regions [57]. Generally, motif analysis methods can be classified into two types: de novo motif discovery that identifies potential new binding motifs for unknown factors appearing in a large fraction of peaks [58]; and motif scanning that estimates and ranks the similarity of supplied DNA sequences against all known canonical motifs within a database [59]. ChIP-seq peaks can also be used in functional enrichment analysis. This analysis binarily labels or quantitatively ranks nearby genes as potential targets and groups them by gene ontology or KEGG pathway

Chromatin-state annotation

Chromatin-state annotation, also called semi-automated genomic annotation (SAGA), classifies all genomic regions by characteristic epigenomic patterns, such as promoters, enhancers, transcribed regions, and repressed regions, using an unsupervised machine-learning approach [63]. Obtained clusters are manually annotated as chromatin states. Typical region-specific analysis (e.g., enhancer analyses [19,64]) narrows down the target genomic regions to be investigated. In contrast, chromatin-state annotation segments the genome and assigns chromatin states to whole-genome regions using a hidden Markov model [65–67] or a dynamic Bayesian network [68]. In this analysis, the biologically optimal number of states is unknown and must be experimentally defined. That is, more abundant states cause difficulty when interpreting obtained clusters. In fact, numerous states may not capture sufficiently distinct epigenetic characters [69]. Thus, up to 15 states may be appropriate. The obtained chromatin states are further extended for various downstream

analyses. For example, ChromDiff [70], EpiCompare [71], and ChromDet [72] combine and cluster derived epigenomic landscapes across multiple cell types to explore tissue or cell typespecific epigenomic regions. A probabilistic clustering approach is also adopted to capture chromatin state dynamics across multiple cell lines [73] or time points [18,74]. Graph-based regularization (GBR) integrates chromatin interaction information for chromatin-state annotation [63]. Generated chromatin state information is then used to interpret individual genetic variations [75,76] and understand epigenetic variation in evolution.

Advanced applications

Because abundant ChIP-seq data are available for several well-studied cell types, it is useful to leverage information from these cell types to infer genome dynamics or to annotate the epigenetic landscape of other cell types with fewer additional experiments. Increasing evidence suggests that epigenetic information is highly correlated with, and can be used to predict, gene expression and chromosomal conformation. In this section, we briefly describe tools for advanced applications of ChIPseq analysis for histone modifications, which are more experimental and theoretical than the tools introduced in section 2.

Gene expression prediction from the epigenome

Various machine learning-based approaches have been developed to quantitatively infer gene expression levels based on the epigenetic information obtained by ChIP-seq experiments. For instance, Karlic et al. applied a linear regression model to histone modification enrichments at promoter sites to predict gene expression in CD4 + T-cells [78]. They utilized nineteen histone modifications and suggested that as few as three promoter site modifications are sufficient to model gene expression [78]. Dong et al. used non-linear models, such as multivariate adaptive regression splines (MARS) and random forests, to map eleven histone modifications and DNase I hypersensitivity in seven human cell lines [79] and successfully predicted gene expression level (Pearson coefficient r = 0.83 with observed data). These models simply consider the epigenetic pattern at promoter sites and do not account for enhancer site information. In contrast, DeepExpression [80] utilizes HiChIP data [81], a high-throughput technique for capturing proteincentric chromosome loops, to consider enhancers and enhancer-promoter interactions. There are also several tools that use convolutional neural networks (CNN) to predict gene expression [82] or differential gene regulation patterns [83]. See reference [82] for a detailed discussion regarding the comparison of these gene expression prediction programs. Considering that the preparation of a single RNA-seq sample requires relatively lower cost compared with that of ChIP-seq samples of multiple histone modifications and HiChIP data, the main purpose of these studies is to elucidate the combinatorial roles of histone modifications in gene regulation, rather than the prediction of gene expression level itself.

Prediction of chromatin interactions from epigenome data

Because recent evidence suggests that single nucleotide polymorphisms (SNPs) in enhancers can cause genetic diseases and cancer [84,85], there is a great demand for genome-wide analysis to characterize the role of enhancers in specific cell lines. However, genomewide pairing of enhancers and target genes is not a trivial task. Indeed, enhancers do not necessarily regulate the nearest genes, and some enhancers are distant from TSSs [86]. While Chromosome Conformation Capture (3C) assays, such as Hi-C [87], HiChIP [81], and ChIA-PET [88], are available to quantify spatial proximity across an entire genome, computational tools for pairing enhancers and target genes keep evolving. Hariprakash and Ferrari classified gene-enhancer pairing tools into four categories [89]: correlation-based, supervised learning-based, regression-based, and score-based. The key differences are "whether multiple enhancers are considered for each gene" and whether multiple epigenetic data are considered for each enhancer/ promoter site". Correlation-based methods estimate the interaction strength for all-by-all enhancer-promoter pairs, while regression-based methods assume that multiple enhancers contribute to a single gene. Supervised learning-based and score-based methods can combine multiple ChIP-seq datasets and other information types for each site (e.g., evolutionary conservation). While these tools focus on enhancerpromoter interactions, there are many other chromatin interactions, such as enhancer-enhancer loops and weak chromatin aggregation via phase separation [90]. In contrast, CITD [91] and DRAGON [92] comprehensively decipher three-dimensional genome organization from epigenetic data using wavelet transformation and potential energy functions, respectively. These statistical approaches aim to find consistent patterns in epigenetic data associated with spatial chromatin contacts and predict them without any previous knowledge of genomic architecture. The limitation of these methods is that genomic interactions are considered as qualitative, rather than quantitative, despite their dynamic nature [93]. It was also reported that the current methods involve a training bias due to sharing information of genomic architecture between training and validation datasets [94]. Nevertheless, because the number of tools is rapidly growing, future methods might achieve sufficient accuracy that identifying enhancer-promoter interactions via 3C-based data will be unnecessary

Data imputation: Reconstruction and denoising ChIP-seq data

One analytical challenge in large-scale ChIP-seq analysis arises from biases and batch effects in ChIP-seq data. Because machine-learning approaches are sensitive to noise in training data, it is unavoidable that some ChIP-seq samples will be identified as moderate quality or rejected as low-quality data (resulting in missing data), especially in cases where multiple laboratories were responsible for data acquisition (e.g., the large consortium project). If biological samples are precious (e.g. primary cells and clinical samples), it might be practically difficult to collect more samples. In this case, "data imputation" methods may be appropriate. These methods utilize many epigenetic data from other closely related cell types for data de-noising or reconstruction. "Data de-noising" aims to improve existing ChIP-seq

sample quality by identifying and removing noise from the data. For example, Coda [95] encodes a generative noise process and recovers signals in ChIPseq data using convolutional neural networks. “Data reconstruction” aims to generate missing ChIP-seq data from the large dataset in silico. ChromImpute [96] is a pioneering tool that trains a regression tree to infer signal from each missing experiment using the ten most correlated cell types. PREDICTD [97] and Avocado [98] leverage tensor decomposition to impute multiple ChIP-seq data simultaneously. Several prediction tools for transcription factor binding sites are also proposed [99–101]. These data imputation approaches are potential computational alternatives to real ChIP-seq experiments, and might open the way to collect epigenomic data for all possible cell types and environmental conditions that are clearly impossible in biology. At the present stage, there are the limitations for the prediction of sample-specific signals that do not correlate with the other samples and for the incorporation of genetic variation [96]. Because ‘a prior expectation of signal’ by the imputation across the genome is informative even when high-quality datasets are available [96], the combined use of observed and imputed data is a practically good strategy. Although this approach is computationally challenging, publicly available high-quality data from diverse cell types (Table 1) encourages to accomplish that.

Single-cell ChIP-seq analysis

Recent evidence suggests many cells types, including normal immune cells, serve an essential accessory function in complex tissues and tumors [102]. To elucidate this cellular heterogeneity and cell fate trajectories in developmental processes, various single-cell assays have been developed [103]. Among them, scChIP-seq enables genome-wide profiling of histone modifications and other chromatin-binding proteins at single-cell resolution from low-input samples. Recently, multiple approaches for single-cell labeling and ChIP-seq library preparation have been developed (Table 2) which use microfluidic systems, Tn5 transposase tagmentation, and ChIP-free strategies.

Microfluidic system-based analysis

The first scChIP-seq method, scDrop-ChIP [104], uses microfluidic systems for cell labeling combined with canonical ChIP methods to generate ~ 800 non-duplicated reads per cell. The more recently developed droplet microfluidic method [105] provides higher resolution, producing ~ 10,000 non-duplicated reads per cell. The limitation of these methods is that the specialized microfluidic devices are not usually available for most laboratories.

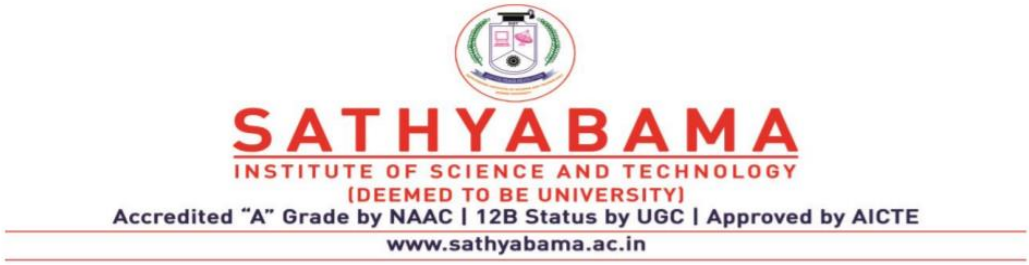
Tagmentation-based analysis

Tagmentation-based library preparation using Tn5 transposase has been widely used for various NGS assays, including ChIP-seq. sc-itChIPseq [106] employs tagmentation for single-cell labeling and library preparation before the canonical ChIP experiment. This method generates ~ 9000 non-duplicated reads per cell. Because the experimental procedure

is similar to the canonical ChIP-seq method, this method is much easier to use than scDrop-ChIP.

ChIP-free methods

Several ChIP-free strategies have been developed for scChIP-seq. Single-cell chromatin immunocleavage sequencing (scChIC-seq) [107] and single-cell uliCUT&RUN [108] are based on the CUT&RUN method [109] that employs MNase and protein A fusion proteins to detect cleaved target sites with a specific antibody. These methods generate ~ 4,100 non-duplicated reads per cell and require several canonical steps for library preparation. However, these methods are limited by low read-mapping rates (~6%). Three similar methods, called CUT& Tag [110], ACT-seq [111], and CoBATCH [112], have been developed. These methods use a Tn5 transposase and protein A fusion protein. During library preparation, the primary antibody is captured by the fusion protein after binding the target protein on chromosomes. Then, Tn5 transposase is activated for tagmentation at the protein binding sites. The advantage of these methods is that protein binding site detection and library preparation are performed simultaneously, which drastically reduces experimental procedures and time. Further, these methods are less subject to technical biases introduced by an immunoprecipitation step. Moreover, these methods show ~ 97% mapping rates and generate ~ 12,000 non-duplicated reads per cell. Thus, this ChIP-free method has potential for high-throughput and highquality scChIP-seq analysis. Finally, chromatin integration labelling followed by sequencing (ChIL-seq) [113] is another ChIP-free method that is based on immunostaining rather than ChIP. The method uses a secondary antibody probe conjugated with dsDNA, which contains a T7 RNA polymerase promoter, an NGS adapter sequence, and a Tn5 binding sequence. After capturing the first antibody, the probe DNA sequence is integrated into the target binding sites by Tn5 transposase. Then, the integrated regions are amplified by in situ transcription, followed by RNA purification and library preparation. The method can be used for single-cell analysis, but likely needs several optimizations to achieve high-throughput sequencing. Additional scChIP-seq methods will be developed in future, such as simultaneous detection of multiple histone modifications and/or other chromatin-binding proteins. These advances will enable to capture colocalization of gene-regulating factors on chromosomes in each cell.



SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – V - Next Generation Sequencing– SBI1606

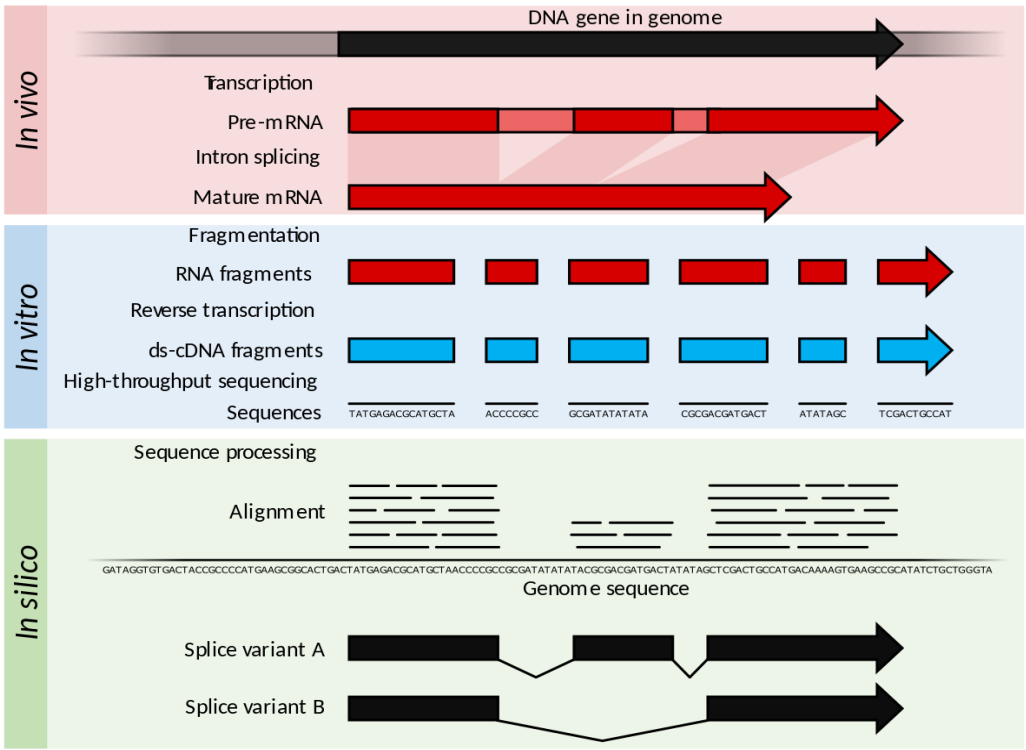
RNA SEQUENCING

RNA-seq is a next generation sequencing (NGS) procedure of the entire transcriptome by which one can measure the expression of several features such as gene expression, allelic expression, and intragenic expression.The number of reads mapped to a given gene or transcript is considered to be the estimate of the expression level of that feature using this technology .Understanding the transcriptome is key if we are to connect the information on our genome with its functional protein expression.RNA-seq can tell us which genes are turned on in a cell, what their level of expression is, and at what times they are activated or shut off.

Microarray technology v/s RNA Seq

RNA-seq is believed to have a wider range of signal detection.The resolution of microarray expression measures cannot go beyond the probe level.RNA-seq can be evaluated at single-base resolution. Moreover, in microarray technology one needs to have knowledge of the target sequences to construct the probe sets. Hence, RNA-seq is more suitable for the discovery of novel transcripts.

Overview of RNA Seq:



DIFFERENTIALLY EXPRESSED GENE

A gene is declared differentially expressed if a difference or change observed in read counts or expression levels/index between two experimental conditions is statistically significant.Such genes are selected based on a combination of expression change threshold and score cutoff, which are usually generated by statistical modeling.The correct identification of differentially expressed genes (DEGs) between specific conditions is a key in the understanding phenotypic variation.

Methods for DGE analysis

1.Parametric

Parametric methods capture all information about the data within the parameters. In these cases, it is possible to predict the value of unknown data from observing

the adopted model and its parameters.Poisson or negative binomial (edgeR & baySeq) uses parametric approach.

2.Non-parametric.

Non-parametric methods can capture more details about the data distribution, i.e., not imposing a rigid model to be fitted.Non-parametric models take into consideration that data distribution cannot be defined from a finite set of parameters, thus the amount of information about the data can increase with its volume.Software tools, such as NOIseq and SAMseq adopt non-parametric methods.

Software for Differential Expression in RNA-seq Data

Several R packages are available for expression analysis, like DEGseq.The Bioconductor software package edgeR has been developed to examine replicated gene count data using an overdispersed Poisson model.The statistical tests based on negative binomial distributions (DESeq, edgeR, and baySeq) had notably good control of false-positive errors with comparable specificity and sensitivity resulted from the tests.

Challenges:

The challenge in analyzing RNA-seq data, particularly in the detection of differential expression, has three primary sources.

1. The inherent problem with the technology.
2. The laboratory or experimental errors causing technical variation across samples.

3. The third and the most important challenge is that current costs of producing RNA-seq data are prohibitive to the generation of many biological replicates, which poses a problem for statistical data analysis.

Case Study:Identification of Differentially Expressed Genes in RNA-seq Data of *Arabidopsis thaliana*: A Compound Distribution Approach

Abstract:

In the present study, the focus is mainly to investigate the differential gene expression analysis for sequence data based on compound distribution model. This approach was applied in RNA-seq count data of *Arabidopsis thaliana* and it has been found that compound Poisson distribution is more appropriate to capture the variability as compared with Poisson distribution.Thus, fitting of appropriate distribution to gene expression data provides statistically sound cutoff values for identifying differentially expressed genes. RNA-seq data of *Arabidopsis thaliana* have been considered for this investigation because of its small size,simplicity,convenience and abundance,susceptibility to T-DNA insertion,short generation time,large number of progeny per plant and small genome of *A. thaliana* make it attractive for molecular genetic analysis.

Why compound distribution approach?

Compound distributions represent a useful way of describing heterogeneity in the distribution of a variable.In the present study, the focus is mainly to investigate the differential gene expression analysis for sequence count data based on compound distribution model as this model is able to capture extra variation.Compound mixture of Poisson–gamma distribution is used. The joint likelihood density function is obtained and the parameters of the model are estimated.

Steps involved:

1. The expression data under the two conditions (hrcC and mock) for different genes are arranged.
2. The difference in read counts is taken over two conditions and is plotted.
3. The positive values are up-regulated gene expression values and the negative values are down-regulated gene expression values.
4. The compound Poisson distribution is fitted to both these values separately, and accordingly the parameters of the distribution are estimated.
5. The goodness-of-fit of the model is tested and the fitted distribution is compared with the single-component Poisson distribution using likelihood ratio test.

Methods:

- It is very important to find statistical distribution to approximate the nature of differential gene expression data,Poisson distribution is most commonly used.

POISSON DISTRIBUTION:

- Poisson distribution occurs when there are events that do not occur as outcomes of a definite number of trials of an experiment but that occur at random points of time and space wherein the interest lies only in the number of occurrences of the event, not in its nonoccurrences.

ADVANTAGES:

- Major advantage is it’s simplicity.

- Has only one parameter.

DISADV:

- It constrains the variance of the modeled variable to be equal to the mean.
- Assumptions of Poisson distribution are too restrictive: it predicts smaller variations that are observed in data.
- Therefore, the resulting statistical test does not control type 1 error (the probability of false discoveries).
- Overdispersion problem was solved in count data by using negative binomial distribution.

NEGATIVE BINOMIAL DISTRIBUTION

- The negative binomial distribution has two parameters, the mean and the dispersion, and hence allows modeling of more general mean–variance relationships.

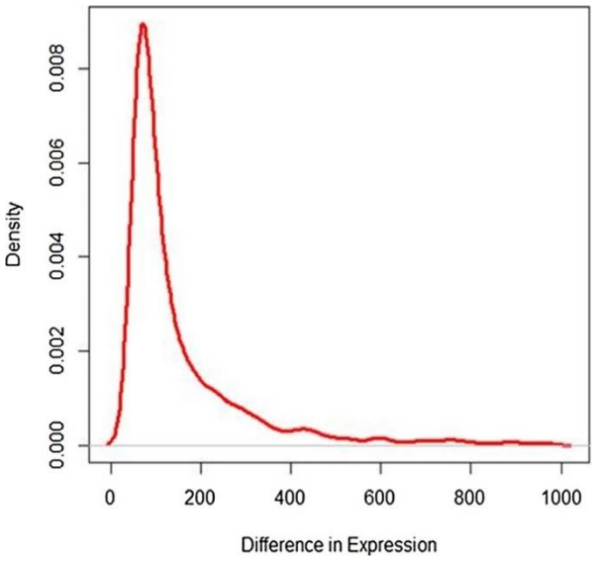
DISADV:

- The number of replicates in the data set of interest is normally too small to estimate both the parameters mean and variance reliably for each gene

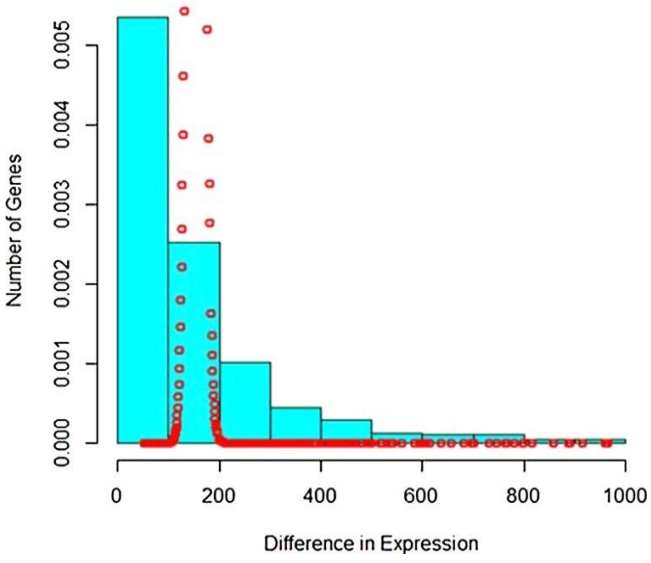
COMPOUND DISTRIBUTION:

- The Poisson parameter is itself a random variable, distributed according to a gamma distribution. The negative binomial distribution is thus here a mixture of a family of Poisson distributions with gamma mixing weights
- Negative binomial as compound Poisson is more capable of capturing the variability as compared with Poisson distribution and hence identified more differentially expressed genes in case of RNA-seq data

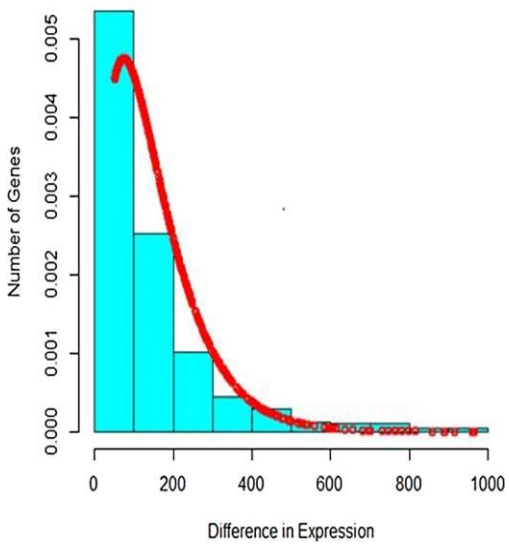
Results:



1.Plot of up-regulated gene expression.



2. Poisson fitting to up-regulated gene expression.



3. Compound Poisson fitting to up-regulated gene expression

Conclusion:

- In case of up-regulated genes, out of 10,483 genes, 9649 genes were identified as differentially expressed based on the probability value cutoff with respect to Poisson distribution and 2081 were identified with compound Poisson distribution.
- Out of a total of 11,607 down-regulated genes, 10,357 were identified as differentially expressed by fitting Poisson distribution, whereas only 1954 were identified with compound Poisson distribution.
- Hence, it can be seen that compound Poisson distribution, which is a mixture of Poisson and gamma, is able to identify the differentially expressed genes more accurately.

Read Mapping

Data analysis of ChIP-seq relies on read mapping. It is important to check the quality of the mapping process. The percentage of mapped reads is a global indicator of the overall sequencing accuracy. This Read Mapping is performed using a reference genome and subsequent identification of signals associated with protein-binding or attachments of modified histones
Most ChIP-seq experiments do not require gapped alignments that consider insertions and deletions (indels) because the sequenced reads do not contain them, unlike exon junctions in RNA-seq analyses.
An important issue concerns the inclusion of multiple mapped. Allowing for multiple mapped reads increases the number of usable reads and the sensitivity of peak detection. However, the number of false positives may also increase. In general, uniquely mapped reads are sufficient to analyze typical TFs, except for in-repeat analyses.
Considering the percentage of mapped reads is important, and desirable rate depends on the species and the read lengths.

Mapping considerations:

Single end reads	**Paired end reads**
Only one end of fragment is known	Both ends of fragment are known
Bowtie can be used	This can done using bowtie2
Up to 50bp	More than 50bp
MACs guess fragment length	MACS now knows mean fragment length

Mapping Tools:

The sequence reads were aligned are in FASTQ or CSFASTQ format from Quality Check. These reads are mapped using any of these tools:

i. Bowtie

ii. Bowtie 2

iii. BWA

BOWTIE

- Among the genome aligners, bowtie is one of a most popular mostly because it can achieve fast alignment.
- Although, the mapping strategy differs between version 1 and 2, the overall pipeline is identical.
- Bowtie uses a "seed and extend" strategy meaning that it will first try to find matches for 5' ends of the reads in the reference genome. In the second step, it will try to extend these matches using dynamic programming.
- In the case of ChIP-Seq analysis, one crucial issue is to control for multi-reads (reads that map to several positions onto the reference genome) that may produce artificial peaks.
- Additionally, flag is set in –m1, that means 1 read only maps to one location (uniquely mapped reads)

BOWTIE 2

- Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy
- -m1 flag is no longer existed here
- Another filtering strategies are required
- Post mapping filtering is also done here
- To check the reads if they are uniquely mapped or not, both read quality and concordancy can be checked using sam tools, flags are shown here,

i. -f2 for concordancy

ii. -q30 for read quality

Difference between BOWTIE AND BOWTIE 2

The chief differences between Bowtie 1 and Bowtie 2 are:

- For reads longer than about 50 bp Bowtie 2 is generally faster, more sensitive, and uses less memory.
- For relatively short reads, less than 50 bp Bowtie 1 is sometimes faster and/or more sensitive.

BWA (Burrows-Wheeler Algorithm)

BWA is a software package for mapping low-divergent sequences based on a Burrows-Wheeler index against a large reference genome, such as the human genome. It consists of three algorithms:

i. BWA-backtrack

ii. BWA-SW

iii. BWA-MEM

The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp.

BWA-MEM and BWA-SW share similar features such as long-read support and split alignment

BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate.

Few BWA parameter

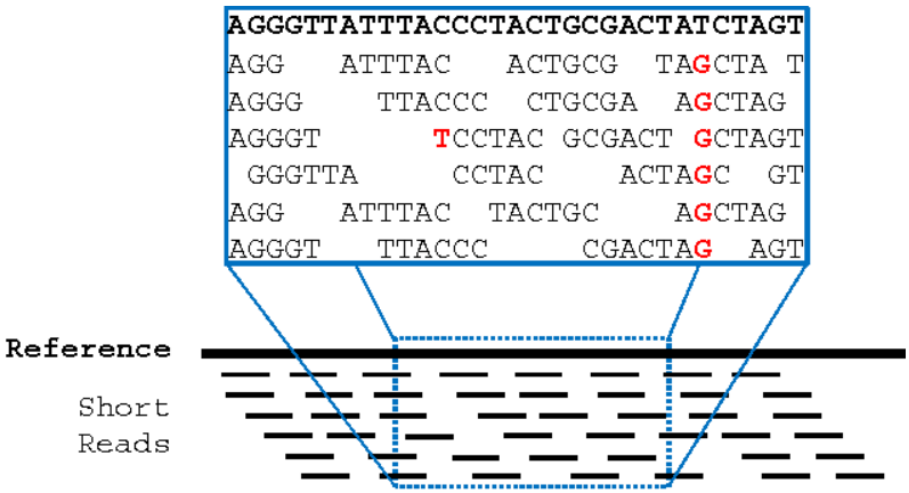
- Command mem works for both single end and paired end reads
- Parameters for long visits are

i. -t : for number of threads we require

ii. -k : minimum seed length (will not match shorter than this length)

iii. -w : band width (longer than this number will not be found)

iv. -T : threshold (regulates BWA that not to generate output lower than the given threshold)



Read Mapping

After mapping

- There are several output formats for map files, such as SAM, BAM, CRAM and tagAlign.
- While the BAM format is the most widely used so far, the more space-Efficient.
- After alignment, reads mapped to the same genomic positions are filtered as redundant reads, and the remaining non-redundant reads are used for analysis.

Peak calling

- The computational analysis is heavily dependent on the detection of “peaks”, regions of the genome where multiple reads align that are indicative of protein binding.
- It is a method used to identify areas in a genome that have been enriched with aligned reads, areas where a protein interacts with DNA.
- numerous peak-calling tools were recently developed for review. However, no tool can achieve 100% accuracy.

- ChIP-seq peaks do not have strand information, it can be estimated from the gene information when focusing on the histone marks that are enriched around TSS (transcription start site).

Peak calling software's

***MACS2** (MACS1.4)*	Most widely used peak caller. Can detect narrow and broad peaks.
***Epic** (SICER)*	Specialised for broad peaks
BayesPeak	R/Bioconductor
Jmosaics	Detects enriched regions jointly from replicates
T-PIC	Shape based
EDD	Detects megabase domain enrichment
GEM	Peak calling and motif discovery for ChIP-seq and ChIP-exo
SPP	Fragment length computation and saturation analysis to determine if read depth is adequate.

MACS2 (Model-based Analysis of ChIP-Seq)

- Most widely used peak caller
- Identifies genome-wide locations of TF binding, histone modification from ChIP-seq.
- Can be used without a control
- Controls eliminate bias due to GC content, mappability or DNA repeats.
- Can call narrow and broad peaks.
- Many settings for optimizing results.
- It uses a method called standard cross correlation, it looks for the regions where fragments are clustered in the genome more than the input fragments are in the same region.
- MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome.
- Results is given out in BedGraph format or WIG format

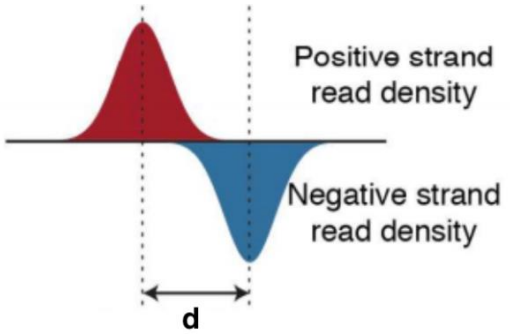
Peak calling in MACS2

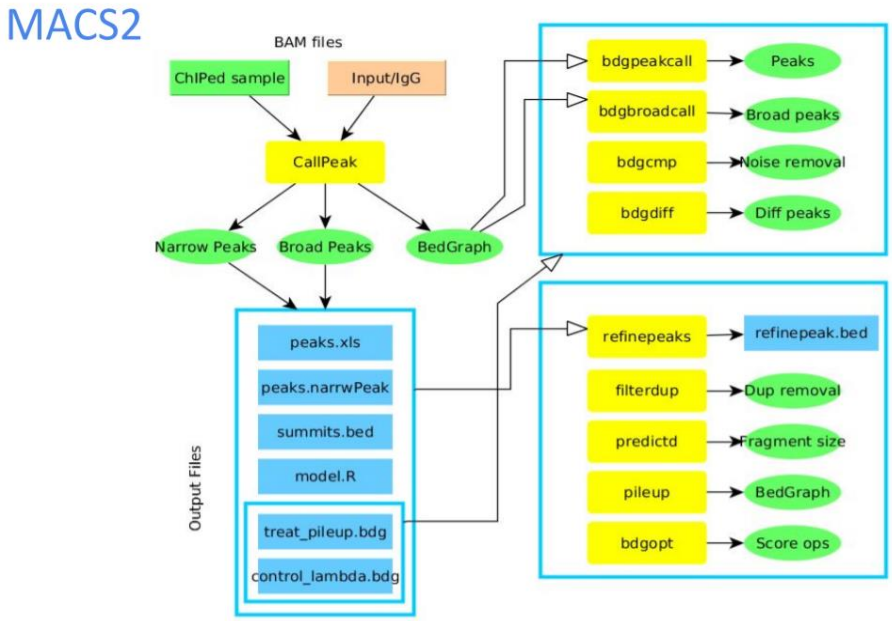
STEP 1: Estimate fragment length d and adjust read position

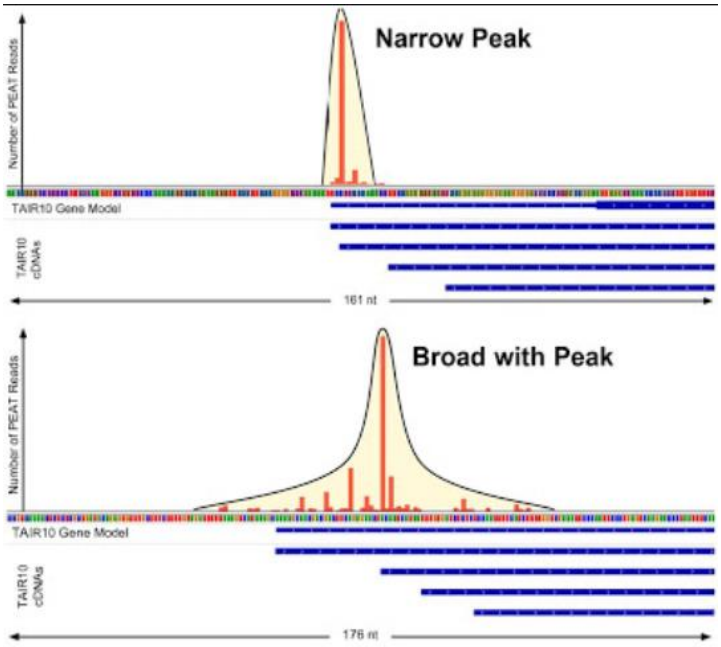
STEP 2: Identify local noise

STEP 3: Identify enriched (peak) regions

STEP 4: Estimate FDR (FDR= negative read/positive reads)







Data Statistical Considerations In Analysis Of Rare Variants:

What is Statistics?

It is the science of learning from data.

- Statistical knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results.
- Helps in making decisions based on data and makes predictions.
- Statistics allows you to understand a subject much more deeply

In short:

•Producing reliable data.

•Analyzing the data appropriately.

•Drawing reasonable conclusions.

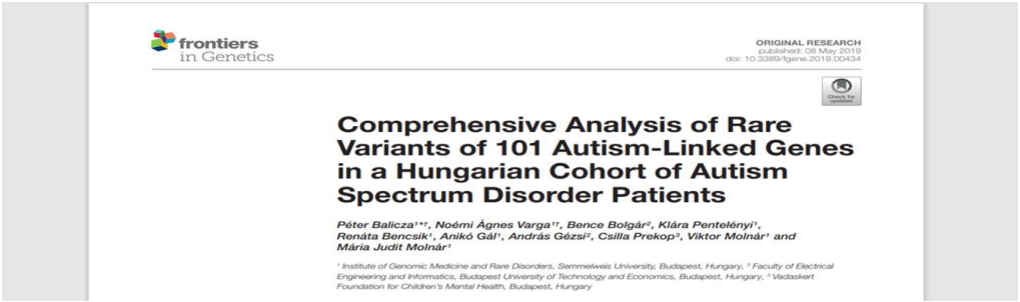
What are Rare variants?

•A rare functional variant is a genetic variant which alters gene function, and which occurs at low frequency in a population. Rare variants may play a significant role in complex disease, as well as some Mendelian conditions.

•Rare variants are alternative forms of a gene that are present with a minor allele frequency (MAF) of less than 1%. (MAF-minor allele frequency)

CaseStudy:

https://www.frontiersin.org/articles/10.3389/fgene.2019.00434/full-



Introduction and Background:

•Autism spectrum disorder (ASD) is genetically and phenotypically heterogeneous. Former genetic studies suggested that both common and rare genetic variants play a role in etiology. In this study, they aimed to analyze rare variants detected by next generation sequencing (NGS) in an autism cohort from Hungary.

•Autism spectrum disorder (ASD) is a neurological and developmental disorder that begins early in childhood and lasts throughout a person's life. It affects how a person acts and interacts with others, communicates, and learns.

•ASD has an estimated heritability of 64–91% , suggesting a strong genetic effect, but the genetic background is highly heterogeneous .

•Common risk variants and rare variants both play a role and mutation types range from single nucleotide variants to large chromosomal aberrations, as well as variations in regulatory DNA elements

Patients:

- Patients Autism spectrum disorder patients were recruited from the Vadaskert Child and Adolescent Psychiatry Hospital and Outpatient Clinic.
- Detailed clinical examinations consisting of a general medical examination and neurological assessment were performed.
- A diagnosis of ASD was made by a qualified psychologist using the ADI-R (Autism Diagnostic Interview-Revised) and ADOS (Autism Diagnostic Observation Schedule).

Materials and Methods:

1. Genetic analysis

•DNA was isolated from peripheral blood samples from all participants using the QIAamp DNA blood kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions

•QIAamp DNA blood kit -For DNA purification from whole blood, plasma, serum, buffy coat, lymphocytes, dried blood spot, body fluids, cultured cells, swabs, and tissue.

•The 101 ASD-associated genes were investigated with NGS, which was performed on a MiSeq (Illumina, San Diego, CA, United States) using the TruSight Autism Rapid Capture Kit (Illumina, San Diego, CA, United States) and the SureSelect QXT Kit (Agilent Technologies, Santa Clara, CA, United States) according to the manufacturer's instructions.

2.Bioinformatical and Statistical analysis:

1. Raw sequences were filtered with Picard tools and quality filtered reads were aligned to the hg19 reference genome with BWA-mem using default parameters

.

2. Variant calling was performed using GATK HaplotypeCaller (version 3.3-0)

3. Variant quality was assessed by GATK, and only variants, which were flagged as PASS (Read depth >10, Mapping quality >40, quality by depth >2) were analyzed.

4. To filter potentially causal single-gene Mendelian variations on a case-by-case level, we used the VariantAnalyzer software developed at the Budapest University of Technology and Economics. This software application annotates SNPs and short INDELs with several types of annotations

5. Finally, mutations were prioritized based on their predicted effects. Exonic frameshifts, stop mutations and canonical splice site variants were considered damaging, whereas the effects of missense mutations were predicted using multiple prediction tools: SIFT ,Polyphen2 , CADD , Radial SVM.

Statistical Considerations Used In Analysis:

For the analysis of rare variants in a multifactorial hypothesis framework on a cohort level, the following methods were used:

1. We tested whether the total number of detected rare missense or loss of function (stop, canonical splice site, and frameshift) variants in a given gene is greater than expected, with the method described by Rao and Nelson (2018). We filtered rare variants, with a MAF cut-off of 5% in the 1000 Genome European dataset, and in our internal exome database of 200 patients.

.P-value was calculated with the associated software: SORVA3 .

1. The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant.

2. For the calculation of rare variant burden, genes were normalized according to genetic intolerance to mutation.

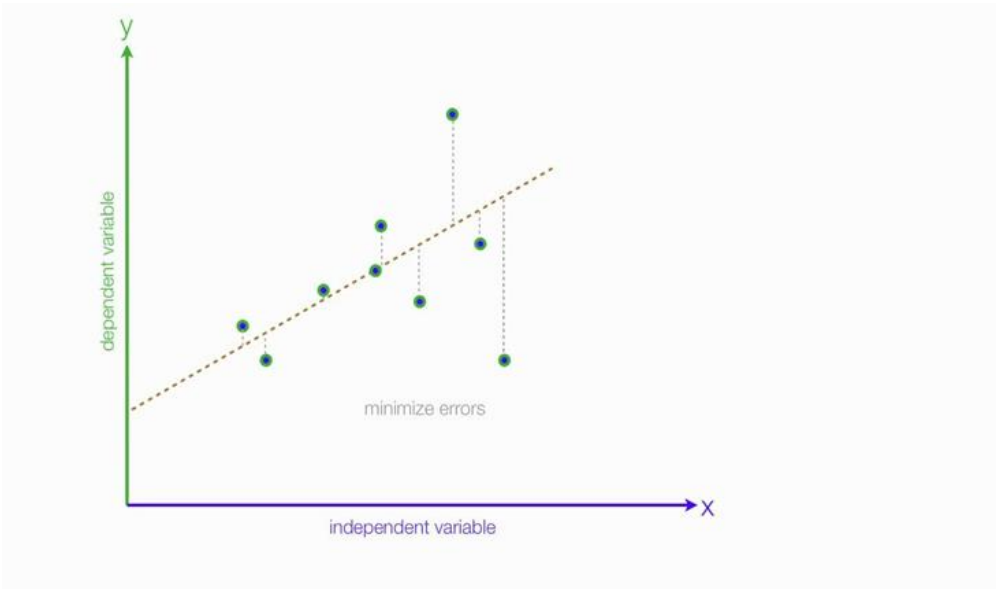
They used the inverse RVIS percentile [1–(RVIS percentile÷100)] to give a weight to every gene .

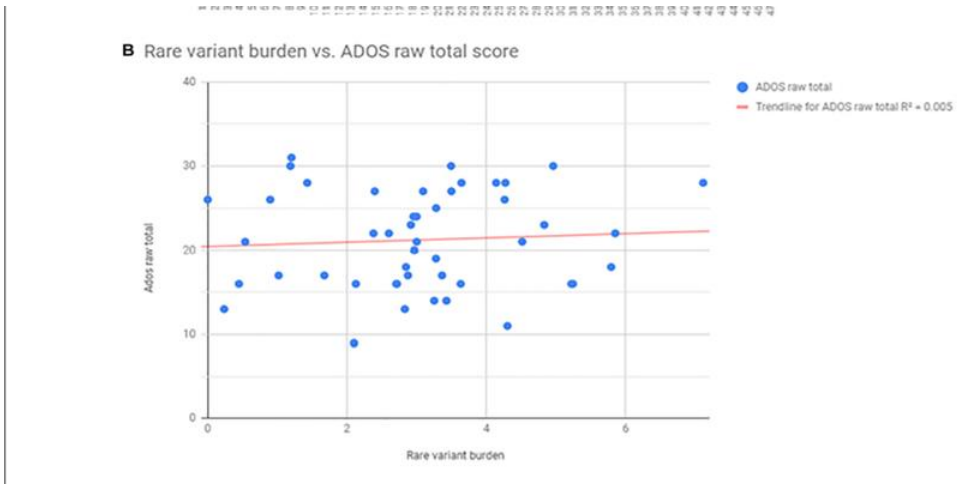
-http://genic-intolerance.org/

- Residual Variation Intolerance Score (RVIS). An RVIS < 0 means that a gene has fewer common functional mutations that expected; an RVIS > 0 indicates that a given gene has a comparatively high frequency of mutations that affect function.
- Linear regression was used then to test for correlation between rare variant burden and autism severity, and rare variant burden vs. minor malformation burden.

Linear regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.





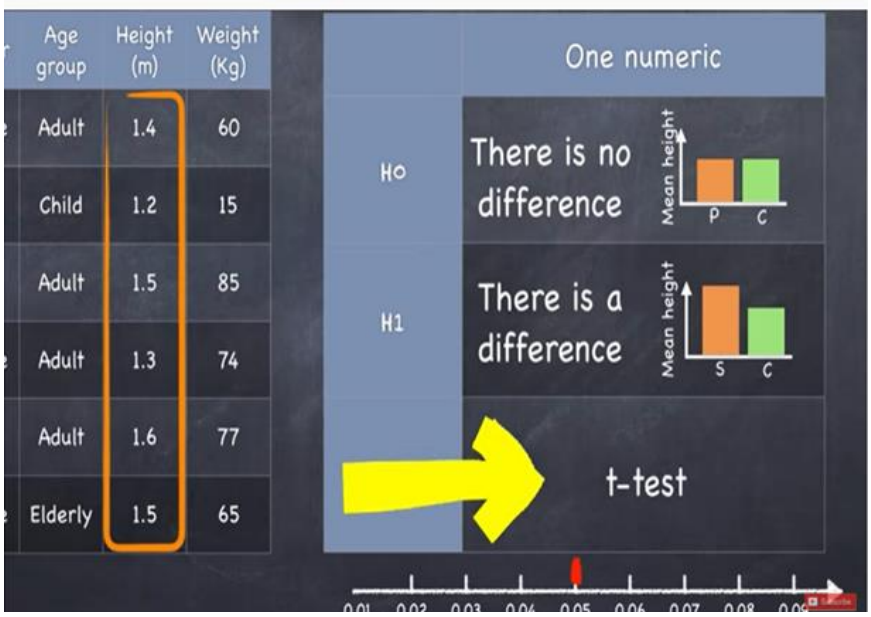
- For comparison of rare variant burden in males versus females, and the number of minor malformations in syndromic versus non-syndromic cases two-tailed T-test was used.

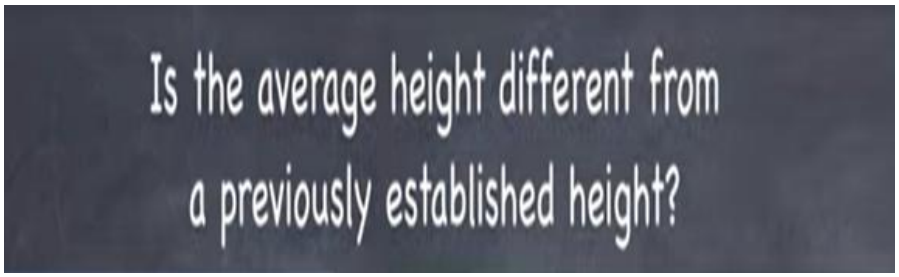
Two tailed t tests:

In statistics, a two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a

certain range of values. It is used in null-hypothesis testing and testing for statistical significance.

Eg for t test:





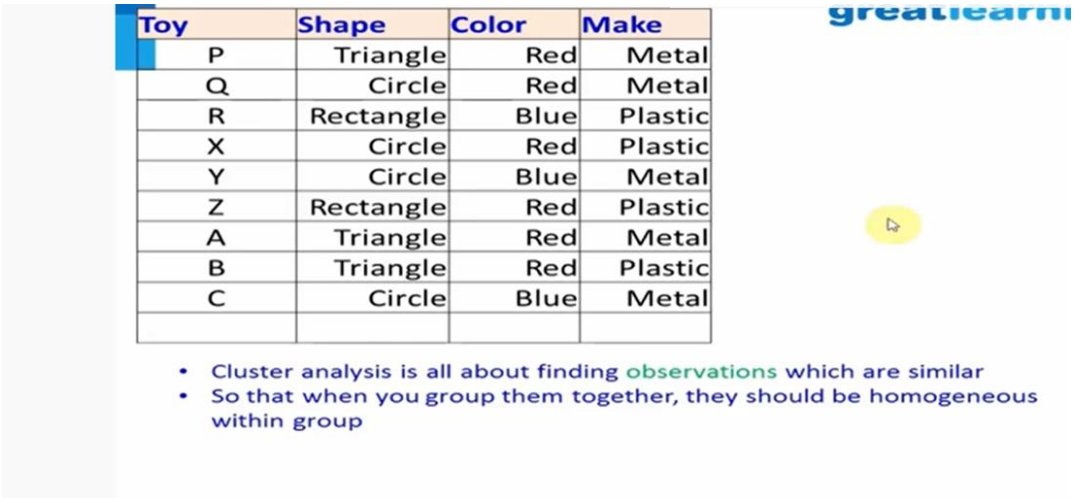
If the p value is less than 0.05 or alpha value, we can reject the null hypothesis and take the alternative hypothesis

3. For the analysis of rare variant association with potential autism subphenotypes they assessed, whether such subphenotypes can be created based solely on the clinical data. They have used clinical questionnaire containing 149 questions about family history, concomitant diseases, drugs, physical examination

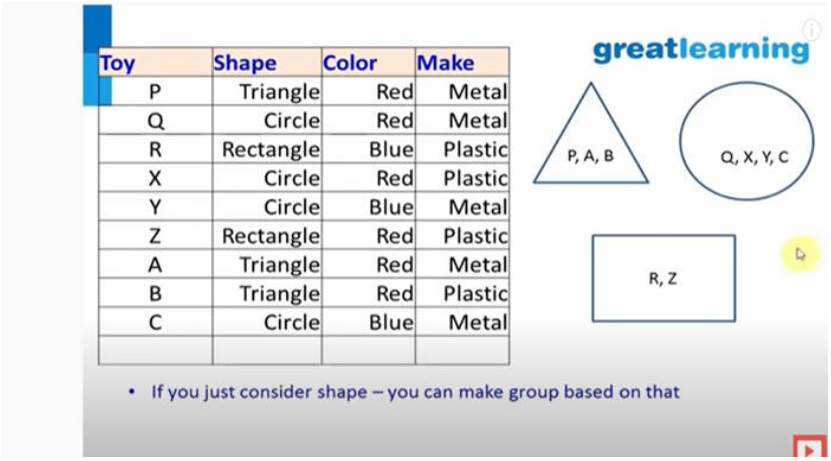
(neurologic and screening of minor malformations), and psychological status for cluster analysis.

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).
- It is also called as subjective segmentation

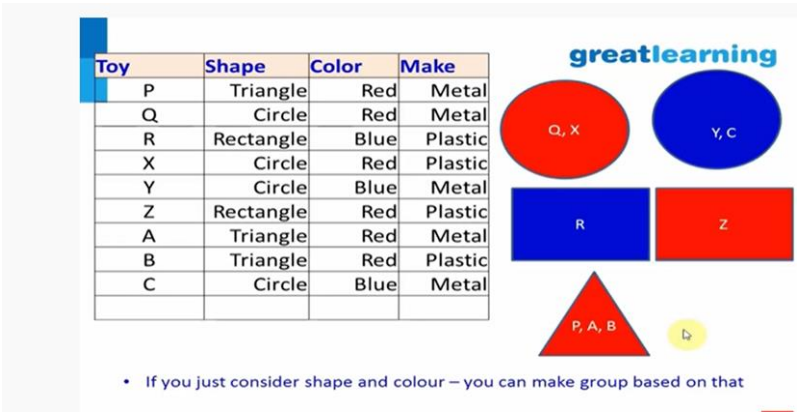
Eg for cluster analysis:



Toy	Shape	Color	Make
P	Triangle	Red	Metal
Q	Circle	Red	Metal
R	Rectangle	Blue	Plastic
X	Circle	Red	Plastic
Y	Circle	Blue	Metal
Z	Rectangle	Red	Plastic
A	Triangle	Red	Metal
B	Triangle	Red	Plastic
C	Circle	Blue	Metal



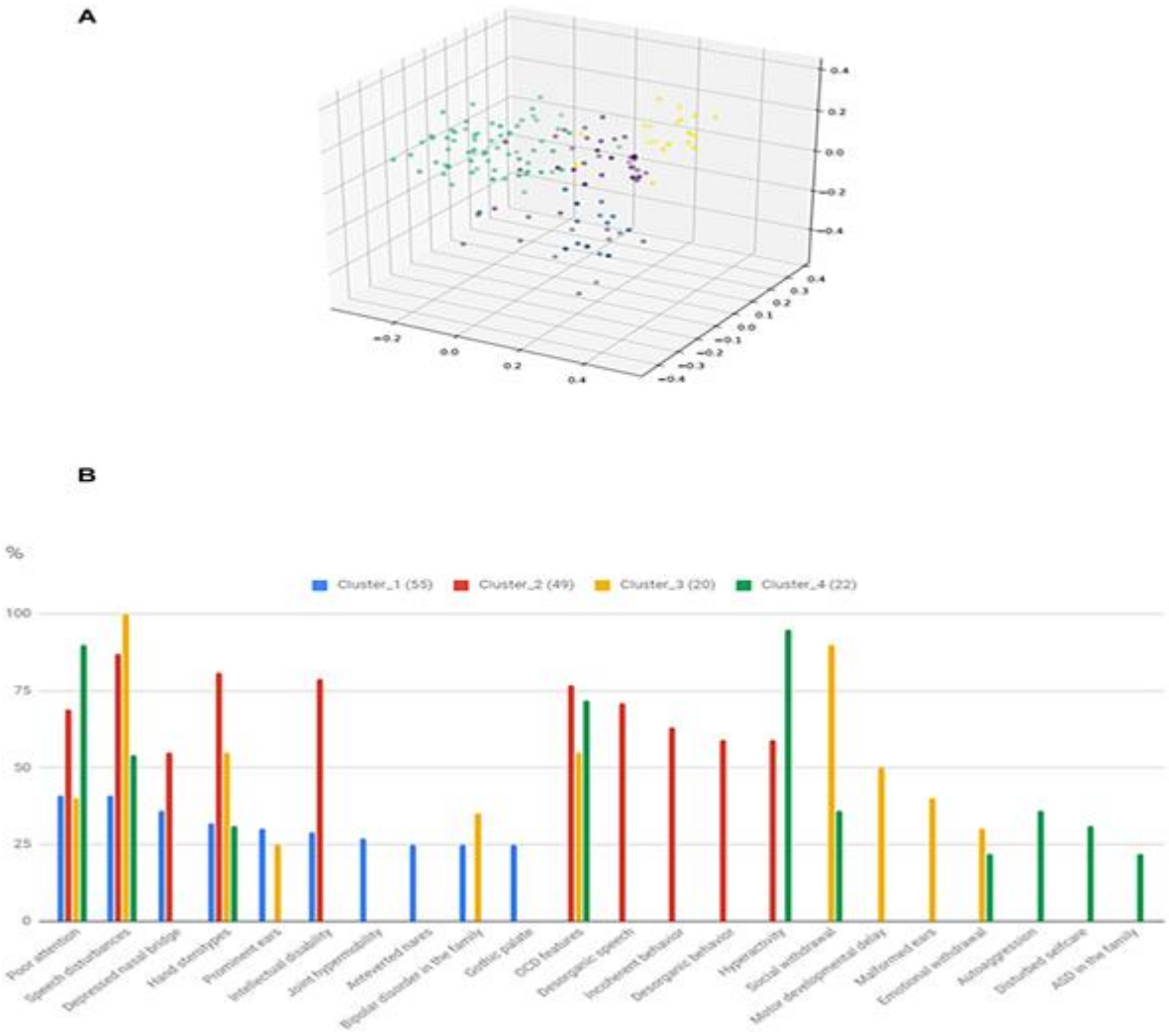
Toy	Shape	Color	Make
P	Triangle	Red	Metal
Q	Circle	Red	Metal
R	Rectangle	Blue	Plastic
X	Circle	Red	Plastic
Y	Circle	Blue	Metal
Z	Rectangle	Red	Plastic
A	Triangle	Red	Metal
B	Triangle	Red	Plastic
C	Circle	Blue	Metal



Toy	Shape	Color	Make
P	Triangle	Red	Metal
Q	Circle	Red	Metal
R	Rectangle	Blue	Plastic
X	Circle	Red	Plastic
Y	Circle	Blue	Metal
Z	Rectangle	Red	Plastic
A	Triangle	Red	Metal
B	Triangle	Red	Plastic
C	Circle	Blue	Metal

- For the phenotypic cluster analysis, given our sample size and the low expected number of clusters, we utilized two kernel-based methods, namely kernel PCA and spectral clustering.
- kernel methods are a class of algorithms for pattern analysis, whose best known member is the support vector machine. The general task of pattern analysis is to find and study general types of relations in datasets.

- Kernel methods have the additional benefit of being non-linear, i.e., able to identify non-linear combinations of clinical variables as relevant features. Any linear model can be turned into a non-linear model by applying the kernel trick to the model



Result of the phenotypic cluster analysis. The three-dimensional figure (A) shows the result of kernel PCA with the identified phenotypic clusters. The histogram (B) represents the relative frequency of the 10 most common features in the given clusters.

- To assess the correlation between the subphenotypes and genetics, we investigated whether detected rare variants of a candidate gene occur more frequently in either of the resulting clusters using ANOVA and pairwise T-tests

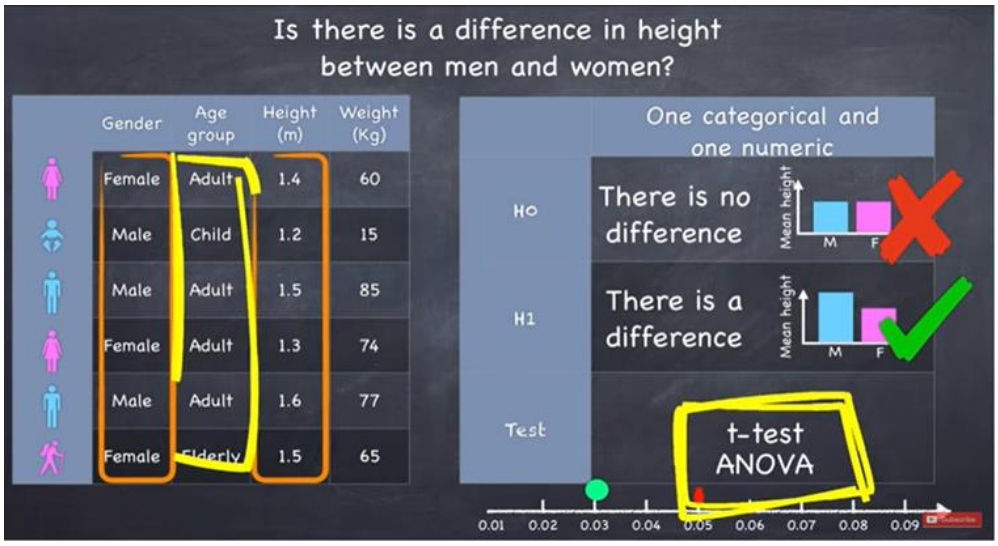
Anova- Analysis of Variance:

- ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.
- Basically, you're testing groups to see if there's a difference between them.
- Examples of when you might want to test different groups:

•A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.

•Students from different colleges take the same exam. You want to see if one college outperforms the other.

- **If there is more than one categorical we use anova, if one categorical and one numerical we can use t tests.**



Sex	**M/F = 129/45 = 2.86/1**
Median age (IQR)	6 (7)/5 (7)
Average ADI	
Reciprocal social interaction	23 ± 5,2
Communication (verbal/non-verbal)	13,95 ± 4,2/18,48 ± 2,5
Repetitive behavior	6,66 ± 2,1
ADOS	21,2 ± 5,7
Comorbid diagnosis	
ADHD (%)	84/174 (48,27%)
Epilepsy (%)	20/174 (11,49%)
Intellectual disability (%)	70/174 (40,22%)
Ethnicity	
Reported Hungarians	170
Reported Romani	2
Other	2
Maternal age at delivery (95% CI)	30,7 ± 1,05
Paternal age at delivery (95% CI)	33,4 ± 1,29
Parent's education level	
College or higher at both parents	34%
College or higher at one parents	27%
High school or lower at both parents	39%

The table describes the characteristics of the cohort. Ethnicity is self-reported. M/F is male to female ratio.

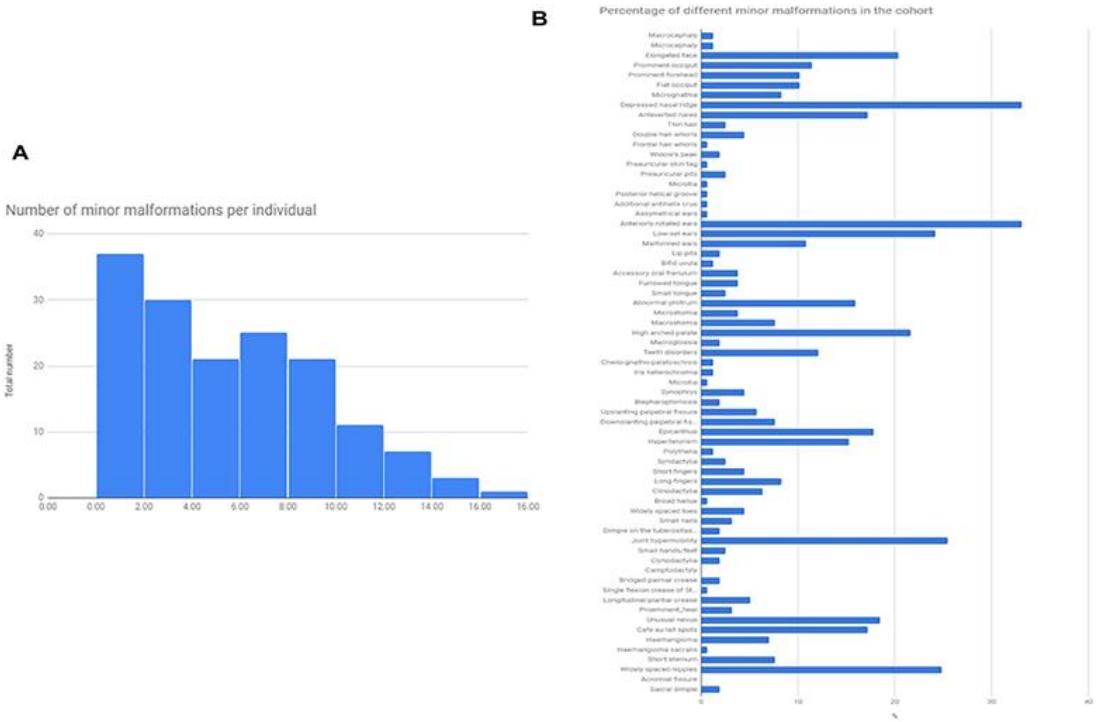


Figure 1. Total number of minor malformations, and the percentage of different minor malformations in the cohort. The Figure represents the total number of minor malformations/given individuals as a histogram (A), and the prevalence of different minor malformations, as a percentage of the total cohort (*N* = 174) (B).

Results and conclusion:

Results:

We have diagnosed 13 molecularly proven syndromic autism cases. Strongest indicators of syndromic autism were intellectual disability, epilepsy or other neurological plus symptoms. Rare variant analysis on a cohort level confirmed the association of five genes with autism (AUTS2, NHS, NSD1, SLC9A9, and VPS13). We found no correlation between rare variant burden and number of minor malformation or autism severity. We identified four phenotypic clusters, but no specific gene was enriched in a given cluster.

Conclusion:

Our study indicates that NGS panel gene sequencing can be useful, where the clinical picture suggests a clinically defined syndromic autism. In this group, targeted panel sequencing may provide reasonable diagnostic yield. Unselected NGS panel screening in the clinic remains controversial, because of uncertain utility, and difficulties of the variant interpretation. However, the detected rare variants may still significantly influence autism risk and subphenotypes in a polygenic model, but to detect the effects of these variants larger cohorts are needed.

WHAT IS GWAS?

GENOME-WIDE ASSOCIATION STUDY (**GWAS**) IS AN APPROACH USED IN GENETICS RESEARCH TO ASSOCIATE SPECIFIC GENETIC VARIATIONS WITH PARTICULAR DISEASES.

THIS METHOD INVOLVES SCANNING THE GENOMES FROM MANY DIFFERENT PEOPLE AND LOOKING FOR GENETIC MARKERS THAT CAN BE USED TO PREDICT THE PRESENCE OF A DISEASE.

ONCE SUCH GENETIC MARKERS ARE IDENTIFIED, THEY CAN BE USED TO UNDERSTAND HOW GENES CONTRIBUTE TO THE DISEASE AND DEVELOP BETTER PREVENTION AND TREATMENT STRATEGIES.

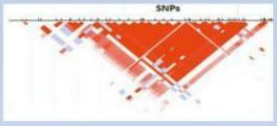
GENERAL WORK FLOW OF GWAS

GWAS workflow

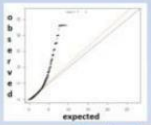
Large cohort (>1000) of cases and controls

Get genome information with SNP arrays

Find deviating from expected haplotypes
visualize SNP-SNP interactions using HapMap



Detection of potential association signals and their fine mapping (e.g. detection of LD, stratification effect)

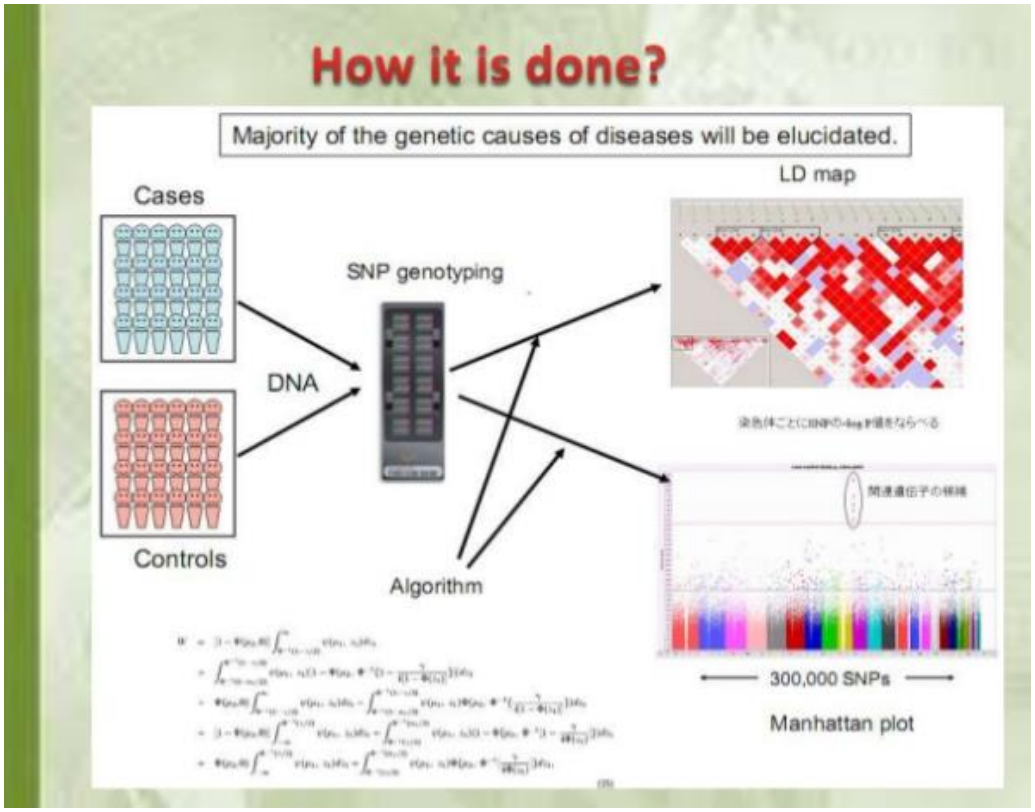


Replication of detected association in new cohot / subset for validation purposes

	AT	AG	Total
cases observed	35	65	100
contorls observed	125	25	100
Totals	160	90	200

Biological / clinical validation

5



FEW TOOLS FOR GWAS

- **GAPIT**

GAPIT (GENOME ASSOCIATION AND PREDICTION INTEGRATED TOOL) IS A TOOL FOR GWAS.THE GAPIT ALGORITHM USES THE MIXED LINEAR MODEL (MLM)

- **GWAS CATALOG**

IS A CATALOG OF PUBLICLY AVAILABLE, MANUALLY CURATED, AND PUBLISHED GENOME-WIDE ASSOCIATION STUDY (GWAS) DATA, CONTAINING OVER 100K SINGLE-NUCLEOTIDE POLYMORPHISMS (SNPS) AND TRAIT ASSOCIATIONS.

- **GARFIELD**

GARFIELD IS AN R TOOL THAT USES GENOME-WIDE ASSOCIATION STUDY (GWAS) DATA TOGETHER WITH ANNOTATIONS TO FIND APPROPRIATE PHENOTYPES.

ADVANTAGE AND DISADVANTAGE OF GWAS

ADVANTAGE

- HAS LESS FALSE POSITIVE RATE
- DISEASE PREDICATION
- DISCOVERY OF NOVEL GENES

DISADVANTAGE

GWAS HAVE MANY LIMITATIONS, SUCH AS THEIR INABILITY TO FULLY EXPLAIN THE GENETIC/FAMILIAL RISK OF COMMON DISEASES; THE INABILITY TO ASSESS RARE GENETIC VARIANTS; THE SMALL EFFECT SIZES OF MOST ASSOCIATIONS; THE DIFFICULTY IN FIGURING OUT TRUE CAUSAL ASSOCIATIONS; AND THE POOR ABILITY OF FINDINGS TO PREDICT DISEASE RISK.

IN THIS REGARD, WE REVIEW BASIC CONCEPTS REGARDING GWAS, THE TECHNOLOGIES USED FOR CAPTURING GENETIC VARIATION, THE MISSING HERITABILITY PROBLEM, THE NEED FOR EFFICIENT STUDY DESIGN ESPECIALLY FOR REPLICATION EFFORTS, REDUCING THE BIAS INTRODUCED INTO A DATASET, AND HOW TO UTILIZE NEW RESOURCES AVAILABLE, WE ALSO LOOK TO WHAT LIES AHEAD FOR THE FIELD, AND THE APPROACHES THAT CAN BE TAKEN TO REALIZE THE FULL POTENTIAL OF GWAS. TO OVERCOME THESE LIMITATIONS WE USE A TECHNOLOGY CALLED NGS-GWAS TECHNOLOGY.

GWAS-NGS TECHNOLOGY

GENOME-WIDE ASSOCIATION STUDIES (GWASS) HAVE BEEN PLAYING AN IMPORTANT ROLE ON HUMAN COMPLEX DISEASES. GENERALLY SPEAKING, GWAS TRIES TO DETECT THE RELATIONSHIP BETWEEN GENOME-WIDE GENETIC VARIANTS AND MEASURABLE TRAITS IN THE POPULATION LEVEL. ALTHOUGH FRUITFUL, GWASS STILL EXIST SOME PROBLEMS, FOR EXAMPLE, THE SO-CALLED MISSING HERITABILITY--SIGNIFICANTLY ASSOCIATED SNPS CAN ONLY EXPLAIN A SMALL PART OF PHENOTYPIC VARIATION. OTHER PROBLEMS INCLUDE THAT, IN SOME TRAITS, SIGNIFICANTLY ASSOCIATED SNPS IN ONE STUDY ARE HARD TO BE REPEATED BY OTHER STUDIES; AND THAT THE FUNCTIONS OF SIGNIFICANTLY ASSOCIATED SNPS ARE OFTEN DIFFICULT TO INTERPRET. HIGH-THROUGHPUT SEQUENCING, ALSO KNOWN AS NEXT-GENERATION SEQUENCING (NGS), COULD BE ONE OF THE MOST PROMISING TECHNOLOGIES TO SOLVE THOSE PROBLEMS BY QUICKLY PRODUCING ACCURATE VARIATIONS IN A HIGH-THROUGHPUT WAY. NGS-BASED GWASS (NGS-GWAS), TO SOME EXTENT, PROVIDE A BETTER SOLUTION COMPARED WITH TRADITIONAL GWASS. WE SYSTEMATICALLY REVIEW THE STRATEGIES AND METHODS FOR NGS-GWASS, PICK OUT THE MOST

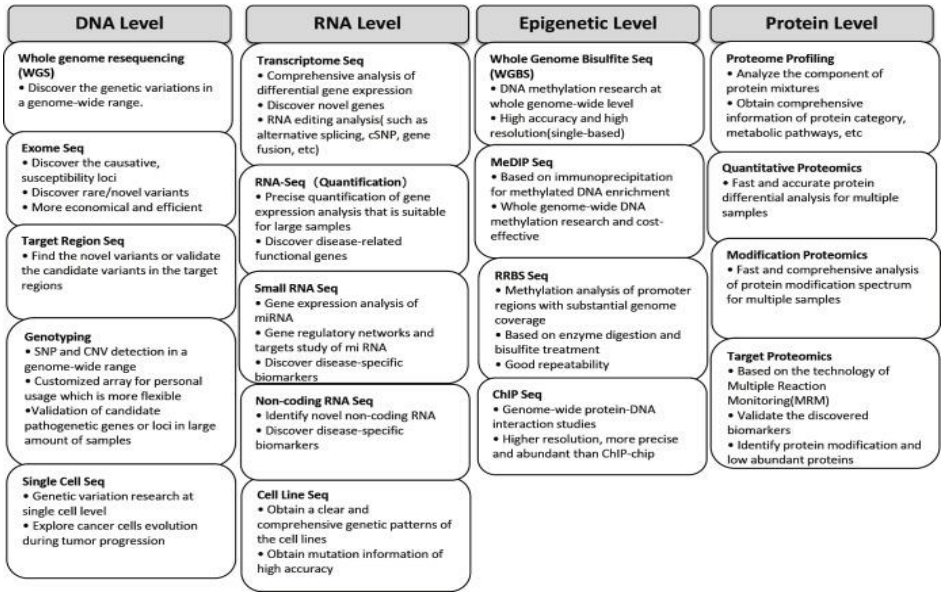
FEASIBLE AND EFFICIENT STRATEGIES AND METHODS FOR NGS-GWASS, AND DISCUSS THEIR APPLICATIONS IN PERSONALIZED MEDICINE.

CASE STUDY ON THE IMPACT OF GWAS-NGS METHOD ON CARDIOVASCULAR DISEASE RESEARCH

ABSTRACT

IN RECENT YEARS, HUNDREDS OF GENE LOCI ASSOCIATED WITH MULTIPLE CARDIOVASCULAR PATHOLOGIES AND TRAITS HAVE BEEN IDENTIFIED THROUGH GWAS-NGS TECHNOLOGY.

THIS SUMMARIZES THE MAIN STRATEGIES OF CV RESEARCH WITH NGS AND GWAS AT THE LEVEL OF GENOMICS, TRANSCRIPTOMICS, EPIGENETICS, AND PROTEOMICS(HOWEVER, BECAUSE OF THE COMPLEXITY OF CVDS, IT IS INSUFFICIENT TO FOCUS ON THE DNA LEVEL ALONE).



INTRODUCTION

- CARDIOVASCULAR DISEASE (CVD) IS A CLASS OF COMPLEX PATHOLOGIES OF THE HEART AND BLOOD VESSELS, INCLUDING CORONARY ARTERY DISEASE (HEART ATTACK), CEREBROVASCULAR DISEASE (STROKE), ELEVATED BLOOD PRESSURE (HYPERTENSION), PERIPHERAL ARTERY DISEASE, RHEUMATIC HEART DISEASE, CONGENITAL HEART DISEASE AND HEART FAILURE.

- It's usually associated with a build-up of fatty deposits inside the arteries (atherosclerosis) and an increased risk of blood clots.

Risk factors: obesity; tobacco smoking; hypertension.

- Sequencing of the entire human genome has exponentially expanded the understanding of genetic contributions to cardiovascular disease. However, these research has demonstrated that static variations of DNA sequence can explain only a fraction of the inherited phenotype and this required sequencing and processing of tremendous amount of data. These data suggest that additional epigenetic and gene expression mechanisms are necessary to explain the expression of CV disease in both experimental and clinical settings

- This has become possible with high-throughput technologies, like NGS. For example, exome-capture and whole-genome sequencing could identify rare and novel genetic variants associated with CVDs.

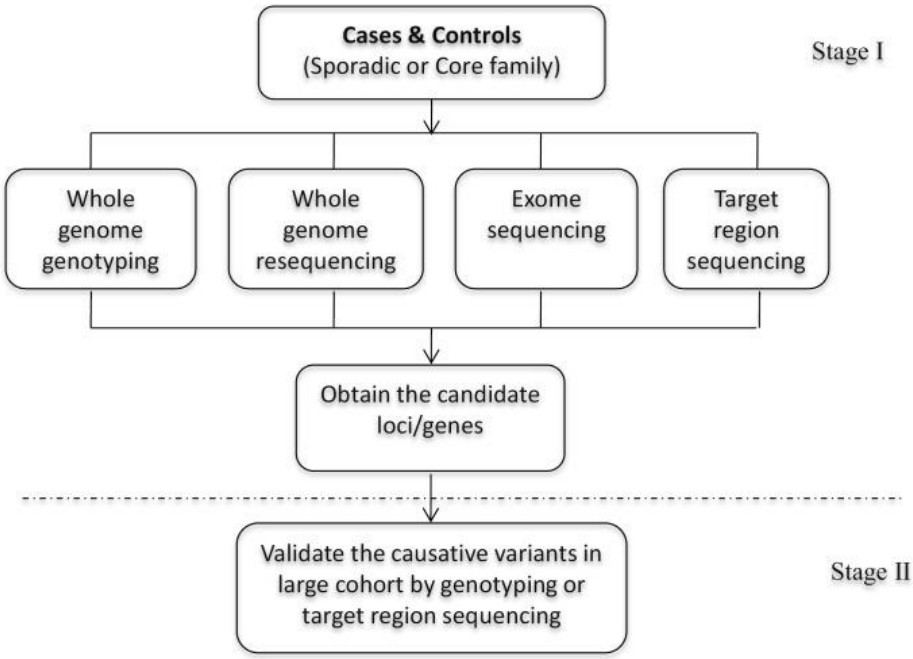
- Eventually, a comprehensive approach like GWAS-NGS will be needed to integrate the accumulated multilevel data.

NGS-GWAS strategy to identify susceptibility/causative genes

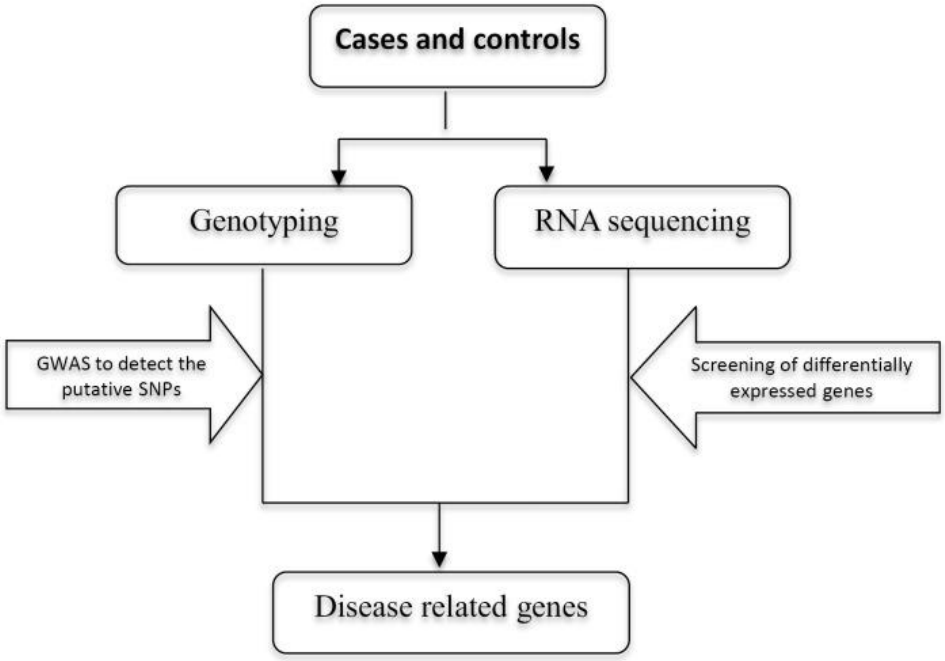
- The wide application of GWAS has led to an enormous boost in the discovery of susceptibility genes for CVDs. Multiple novel genetic loci have been identified in common cardiovascular conditions, including myocardial infarction, hypertension, heart failure, stroke and hyperlipidemia. Up to now, 26 risk loci have been identified by GWAS to be associated with coronary artery diseases. However, only a small fraction of the heritable risk for CVDs can be explained by the variants identified by current GWAS.

- GWAS is based on the common disease-common variant hypothesis, and could provide information on how common genetic variability confers risk for the common diseases while NGS could pinpoint novel genes that contain mutations underlying the phenotype.

- THE STRATEGY OF NGS-GWAS COMBINES NEXT-GENERATION SEQUENCING AND GENOTYPING TO UNCOVER NOVEL CAUSATIVE GENETIC VARIANTS OF COMPLEX DISEASES. COMPARED WITH TRADITIONAL GWAS, IT CAN PROVIDE MORE DETAILED INFORMATION, INCLUDING NOT ONLY COMMON SNPS, BUT ALSO RARE VARIANTS.
- THE IDENTIFIED SUSCEPTIBILITY LOCI WILL BE VERIFIED IN MOLECULAR AND PHYSIOLOGICAL STUDIES TO DETERMINE THE MECHANISMS THROUGH WHICH THESE LOCI CONFER SUSCEPTIBILITY



COMBINING CHIP-BASED GWAS AND RNA SEQUENCING TO IDENTIFY DISEASE-RELATED GENES



CONCLUSION

CONSIDERABLE PROGRESS HAS BEEN MADE IN THE FIELD OF GENOME RESEARCH RELATED TO CVD AND HUNDREDS OF LOCI ASSOCIATED WITH CARDIOVASCULAR PATHOLOGIES HAVE BEEN IDENTIFIED.

GWAS IS BASED ON THE COMMON DISEASE-COMMON VARIANT HYPOTHESIS, AND COULD PROVIDE INFORMATION ON HOW COMMON GENETIC VARIABILITY CONFERS RISK FOR THE COMMON DISEASES WHILE NGS COULD PINPOINT NOVEL GENES THAT CONTAIN MUTATIONS UNDERLYING THE PHENOTYPE.

THE INTEGRATION OF MULTI-OMICS DATA WILL ENABLE CLEARER UNDERSTANDING OF DISEASE-ASSOCIATED LOCI.

HENCE, WITH THE NGS-GWAS STRATEGY 100'S LOCI ASSOCIATED WITH CV WERE IDENTIFIED EFFICIENTLY.

Chip sequencing Data Analysis

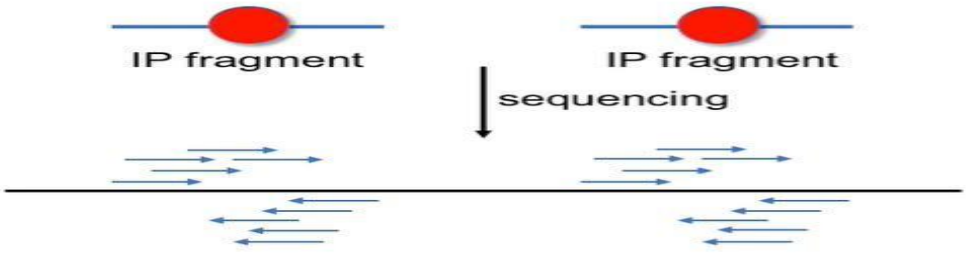
- The ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

- It can be used to map global binding sites precisely for any protein of interest. Previously, Chip-on-chip was the most common technique utilized to study these protein–DNA relations

Steps Involved in the Data Analysis

1. Mapping the reads back to reference genome
2. Background Estimation
3. Peak calling
4. Peak annotation
5. Denovo motif Analysis

1. Mapping the reads back to reference genome

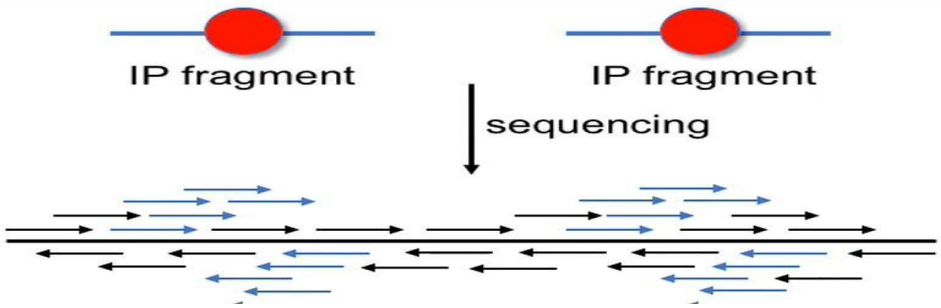


1) From the Immuno precipitant DNA fragments, the reads covering those fragments are obtained as shown by mapping them to the reference genome(Blue).

2) In this step our goal is to identify, for each short read in the dataset, all the locations in a reference genome that show perfect or near perfect matches to the read.

3) Likewise we also get the background datasets or the other DNA present in the genome or Noise which are to be cleared in the further steps.

2. Background Estimation

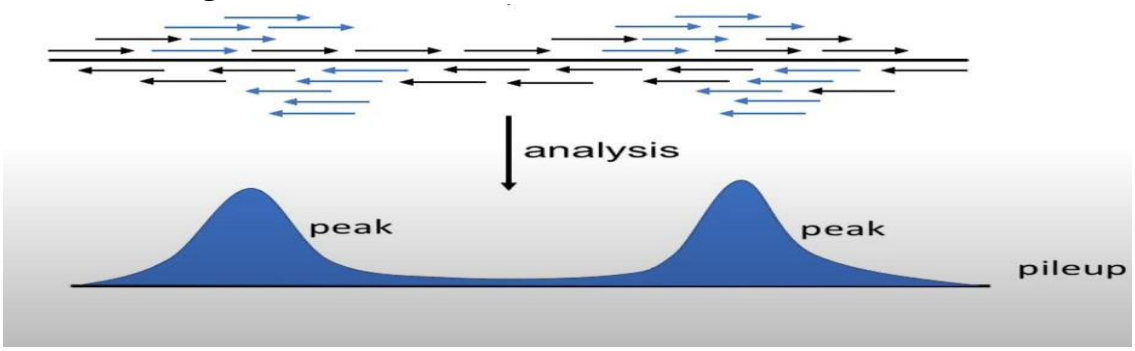


1) In any ChIP-seq datasets, a considerable fraction of the reads may not have originated from these ChIP fragments (Black) .

2) For example the antibody might target proteins other than the one studied, therefore capturing nonspecific fragments. Other factors that may induce such extraneous reads include library contamination, PCR amplification selection, linker/adapter contamination, and image processing errors.

3) We call a read a true signal read if it falls into the called peak regions Otherwise, we call it a background read .

3. Peak Calling



1)The most critical task in the ChIP-seq data analysis pipeline. This is to identify the ChIP signal enriched genomic regions. In other words**, where did the TF bind?**

2) Plotting this to find out the how many reads are covering each genomic position, we find the peaks in the curve where the coverage is higher than other places in the background.

3) These peaks correspond to the position of the IP fragments. This crucial step in the data analysis is called as the **"peak calling".**

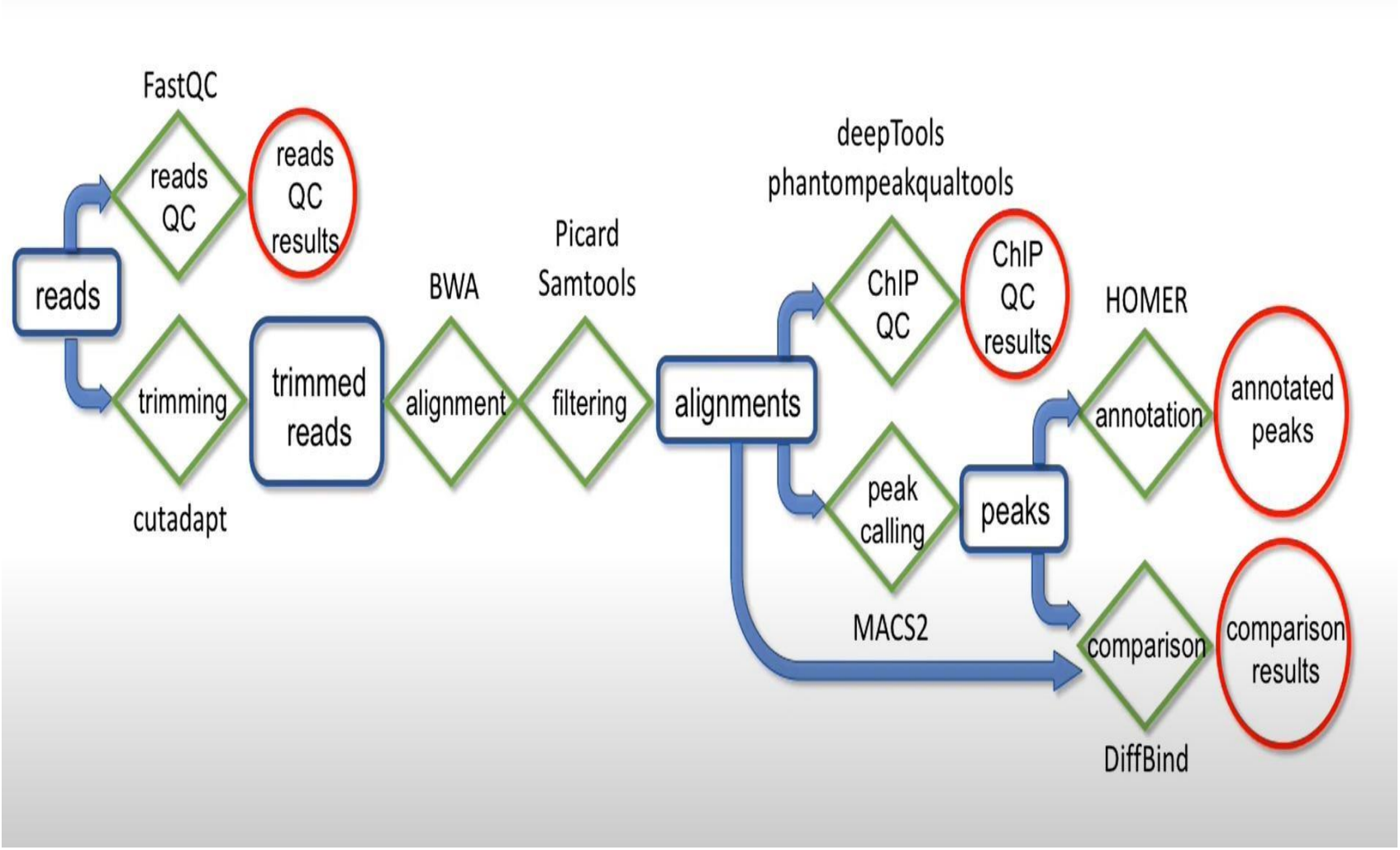
4. Peak Annotation

1) After we obtain a list of peak coordinates, it is important to study the biological implications of the protein–DNA bindings.
2) The number of peaks annotates the quality of the Chip sequencing process.

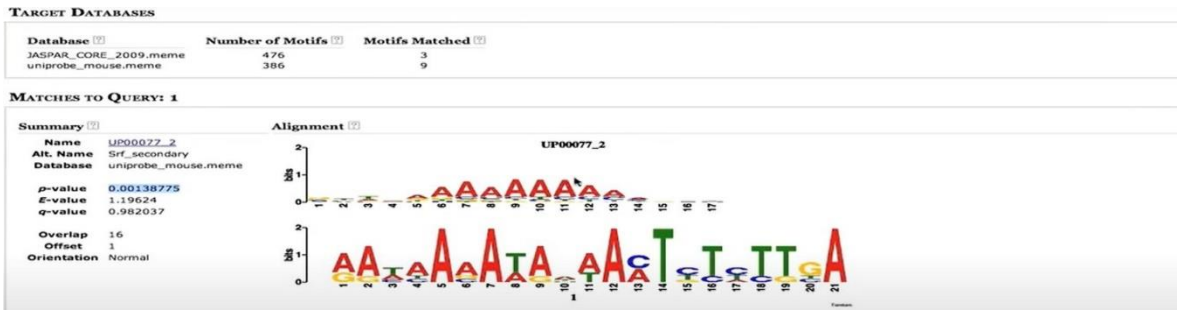
Good – More peaks.

Bad – Less peaks and forms blocks in case of complete failure.

3) The peak calling identifies the binding sites of the proteins of interest.
4) We can identify what genes are near those peaks that are potentially effected by the protein of interest. And this process is called as **"Peak Annotation"**.

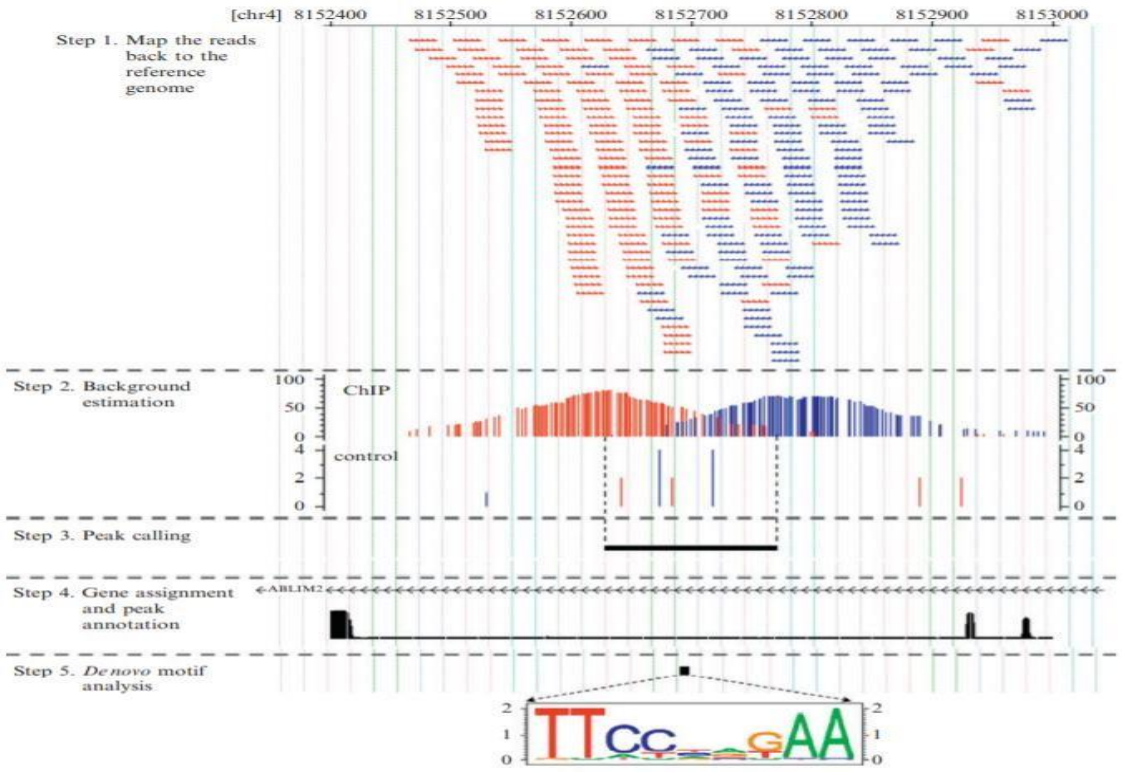

FastQC
reads QC
reads QC results
reads
trimming
cutadapt
trimmed reads
BWA
alignment
Picard
Samtools
filtering
alignments
deepTools
phantompeakqualtools
ChIP QC
ChIP QC results
peak calling
MACS2
peaks
HOMER
annotation
annotated peaks
comparison
DiffBind
comparison results

5. Denovo motif Analysis



1. Another important task in the analysis of the predicted peak regions is de novo motif discovery. In some studies, the exact sequence to which the TF binds is known, or even better, a set of validated binding sites is available. However, if this information is not available, we will need to recover the binding motifs from the peak sequences as well as from their orthologous sequences.
2. Show above is a software called TOMTOM where a specific motif can be given as input in a text format and it matches the motifs to the selected databases and gives out the similar motif results.
3. Motif occupancy and enrichment in peak regions and motif conservation scores offer additional means for assessments.

An Overview



Metagenomics

Metagenomics is the study of metagenome, genetics material, recovered directly from environmental sample such as soil, water , organisms,etc.The term metagenomics first used by Jo Handelsman, Jon Clarly, Robert M. Goodman and first appeared in publication in 1998.

Metagenomics is based on the genomics analysis of microbial DNA directly from the communities present in samples.Metagenomics can unlock the massive uncultured microbial diversity present in the environment for new molecule for therapeutic and biotechnological application.

The science of metagenomics, only a few years old, will make it possible to investigate microbes in their natural environments, the complex communities in which they normally live.

Metagenomics defined as "the genomics analysis of microorganism by direct extraction and cloning DNA from a collection of microorganism."Metagenomics technology – genomics on a large scale will probably lead to great advances in medicine, agriculture, energy production and bioremediation.

HISTORICAL EVENTS IN METAGENOMICS

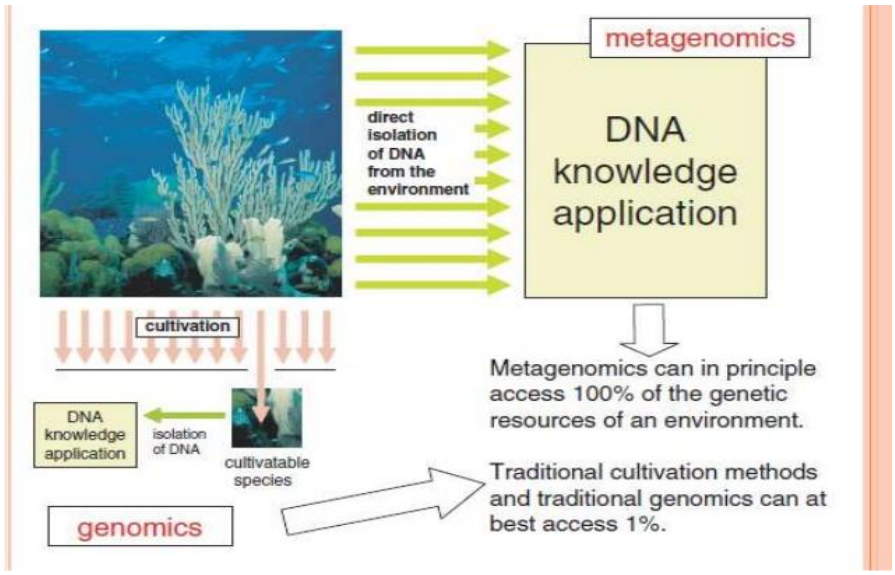
- In 1985 Pace and coworker introduced the idea a cloning DNA directly from environmental samples.
- In 1991 Schmidt and coworker cloning of DNA from Picoplankton in a phase vector subsequent 16S rRNA gene sequence analyses.
- In 1995, Healy reported first successful function driven metagenomics library was screened and termed that Zoolibraies.
- In 2002, Mya Breitbart and Forest Rohwer, used shotgun sequencing to show that 200 liters of seawater contain over 5000 different viruses.

Why metagenomics??

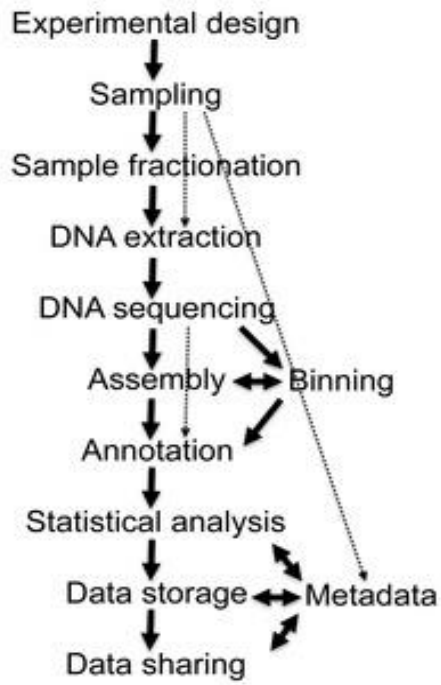
Science of metagenomics make it possible to investigate resource for the development of novel genes, enzymes and chemical compounds for use in biotechnology.Microbes, as communities, are key players in maintaining environmental stability.

Investigate microbes in their natural environment, the complex communities in which they normally live in.

High-throughput gene-level studies of communities.



Steps in Metagenomics



Sampling and Processing

Sample processing is the first and most crucial step in metagenomics.

DNA extracted should be representative of all cells present in the sample and sufficient amounts of high quality nucleic acids must be obtained for subsequent library production and sequencing.

Sample fractionation steps should be checked to ensure that sufficient enrichment of the target is achieved and that minimal contamination of non-target material occurs.

Physical separation and isolation of cells from the samples might also be important to maximize DNA yield or avoid co-extraction of enzymatic inhibitors that might interfere with subsequent processing.

Some type of sample such as biopsies or ground water often yield very small amounts of DNA but in library production for most sequencing technologies require high amounts of DNA (ng or µg), and hence amplification of starting material might be required.

Multiple displacement amplification (MDA) using random hexamers and phage phi29 polymerase is one option employed to increase DNA yields, this method has been widely used in single-cell genomics and to a certain extent in metagenomics.

Types of metagenomics

There are two basic types of Metagenomics studies

I. Sequence-based Metagenomics - involves sequencing and analysis of DNA from environmental samples.

II. Function-based Metagenomics - involves screening for a particular function or activity.

Sequence-based metagenomics studies can be used to assemble genomes, identify genes, find complete metabolic pathways, and compare organisms of different communities

Sequence-based metagenomics can also be used to establish the degree of diversity and the number of different bacterial species existing in a particular sample.

Functional metagenomics involves isolating DNA from microbial communities to study the functions of encoded proteins. It involves cloning DNA fragments, expressing genes in a surrogate host, and screening for enzymatic activities.

DNA SEQUENCING

DNA sequencing is one of the most important platforms for the study of biological systems today.

A. Next generation DNA sequencing

I. 454 life sciences or pyrosequencing
II. Solexa/Illumina
III. Sequencing by ligation (SOLiD technology)
IV. Ion Torrent

Pyrosequencing

- Pyrosequencing is based on the sequencing-by-synthesis principle
- Pyrosequencing has the potential advantages of accuracy, flexibility, parallel processing, and can be easily automated.

- Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA), in which the sequencing is performed by detecting the nucleotide incorporated by a DNA polymerase.

Illumina/Solex

- Immobilizes random DNA fragments on a surface and then performs solid-surface PCR amplification, resulting in clusters of identical DNA fragments.
- Some of the datasets will show the bad errors at the tail ends of reads , we can remove the errors by clipping the reads by using aligners.
- Important factor to consider is run time.

SOLiD technology

- Extensively used, for example, in genome resequencing.
- SOLiD advantage is it provides lowest error rate of any current NGS sequencing technology, however it does not achieve reliable read length beyond 50 nucleotides.
- This will limit its applicability for direct gene annotation of unassembled reads or for assembly of large contigs.

Ion Torrent

- Ion Personal Genome Machine (PGM) based on the principle that protons released during DNA polymerization can detect nucleotide incorporation.
- This system promises read lengths of > 100 bp and throughput on the order of magnitude of the 454/Roche sequencing systems.

ASSEMBLY

Two strategies can be employed for metagenomics samples: reference-based assembly (co-assembly) and de novo assembly.

Reference-based assembly can be done with software packages such as Newbler (Roche), AMOS ,or MIRA. These software packages include algorithms that are fast and memory-efficient.

Reference-based assembly works well, if the metagenomic dataset contains sequences where closely related reference genomes are available.

De novo assembly typically requires larger computational resources. Thus, a whole class of assembly tools based on the de Bruijn graphs was specifically created to handle very large amounts of data.

Machine requirements for the de Bruijn assemblers Velvet or SOAP.

BINNING

Binning is the process of grouping reads or contigs into individual genomes and assigning the group to specific species, subspecies or genus.

More innovative binning approaches include co-abundance gene segregation across a series of metagenomic sample thus facilating the assembly of microbial genomes without the need for reference sequences.

Important considerations for using any binning algorithm are the type of input data available and the existence of a suitable training dataset.

Binning methods can be characterized in two different ways depending on information contained within a given DNA sequence.

1.Composition based binning

2. Similarity or homology based binning

Composition based binning is based on the observation that individual genomes have a unique distribution of k-mer sequence is known as genomic signatures.

Compositional based binning algorithms include phylopythia, successor phylopythiaS, S-GSOM, PCAHIER, TACAO, TETRA, ESOM and ClaMS.

Similarity based binning refer to the process of using alignment algorithms such as BLAST or profile hidden markov models (pHMMs) to obtain similarity information about specific sequences/ genes from publically available databases.

Similarity based binning algorithms include IMG/M, MG-RAST, MEGAN, CARMA, Sort-ITEMS and Metaphyler.

ANNOTATION

Annotation is the process of assigning functional, positional, and species of-origin information to the genes in a database.

Annotation of metagenome is specifically designed to work with mixtures of genomes and contig of varying length.

Steps in Annotation

a) Trimming of low quality reads
b) Masking of low complexity reads-performed using tool such as DUST.
c) De-replication step
d) Screening

Storage and sharing of data

NCBI is mandated to store all metagenomic data, however, the sheer volume of data being generated means there is an urgent need for appropriate ways of storing vast amounts of sequences.

Tools such as IMG/MER, CAMERA, MGRAST, and EBI metagenomics (which also incorporates QIIME) provide an integrated environment for analysis, management, storage, and sharing of metagenome projects.

A suite of standard languages for metadata is currently provided by the Minimum Information about any (x) Sequence checklists (MIxS)

Applications of metagenomics

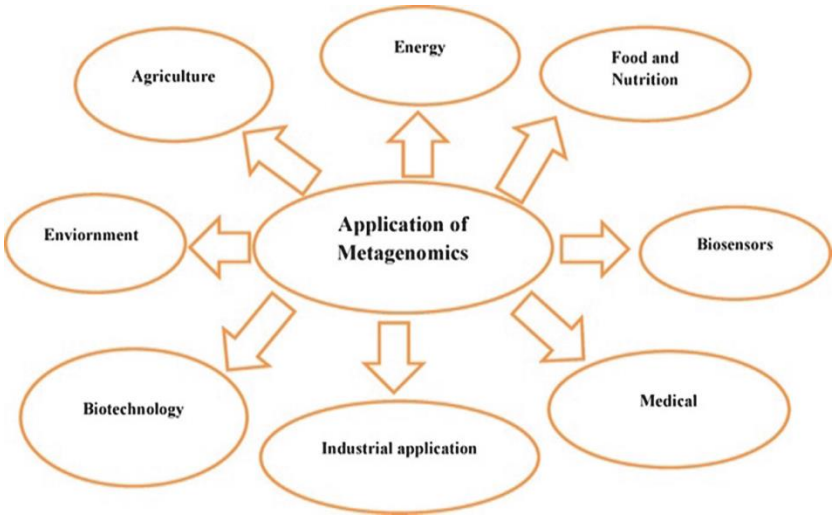
Metagenomics can improve strategies for monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments.

Recent progress in mining the rich genetic resource of nonculturable microbes has led to the discovery of new gene, enzymes and natural products. The impact of metagenomics is witnessed in the development of commodity and fine chemicals, agrochemicals and pharmaceuticals where the benefit of enzyme catalyzed chiral synthesis is increasingly recognized.

Metagenomics libraries are, indeed, an essential tool for the discovery of new enzymatic activities, facilitating genetic tracking for all biotechnological applications of interest for the future.

Metagenomics sequencing is being used to characterize the microbial communities.

Functional metagenomics strategies are being used to explore the interactions between plants and microbes through cultivation-independent study of the microbial communities.



Limitations

- To much data.
- Most gene are not identifiable
- Contamination, chimeric clone sequences
- Extraction problem
- Requires proteomics or expression studies to demonstrate phenotypic characteristics

- Need a standard method for annotating genomes
- Can only progress as library technology progresses, including sequencing technology.
- Requires high throughput instrumentation not readily available to most institutions.

Case study

Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases.

Next generation sequencing (NGS) methods are powerful tools with the potential for comprehensive and unbiased detection of pathogens in clinical samples.

The use of this new technology for the diagnosis of suspected infectious encephalitis, and discuss the feasibility for introduction of NGS methods as a frontline diagnostic test.

The review identified 25 articles reporting 44 case reports of patients with suspected encephalitis for whom NGS was used as a diagnostic tool.

Hundreds of pathogens have been associated with encephalitis, with the most frequently identified including Herpes simplex virus (HSV), Varicella zoster virus (VZV), enteroviruses, Measles morbillivirus, Mumps virus, Japanese encephalitis virus (JEV), influenza viruses, adenoviruses and Mycoplasma pneumoniae.

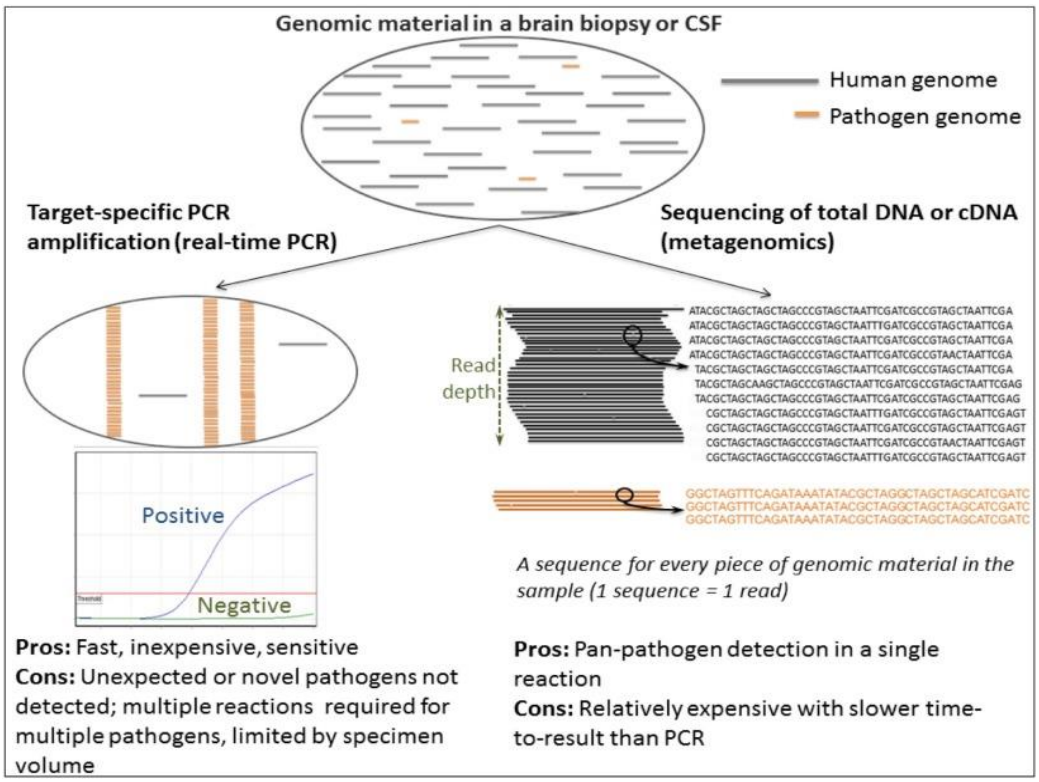
The main alternative etiology to infection is immune mediated, for which management includes immune suppression.

Diagnostics for encephalitis and the role of modern technologies

A laboratory will perform targeted tests for a disease. These are largely confined to specific polymerase chain reaction (PCR) or serological assays.

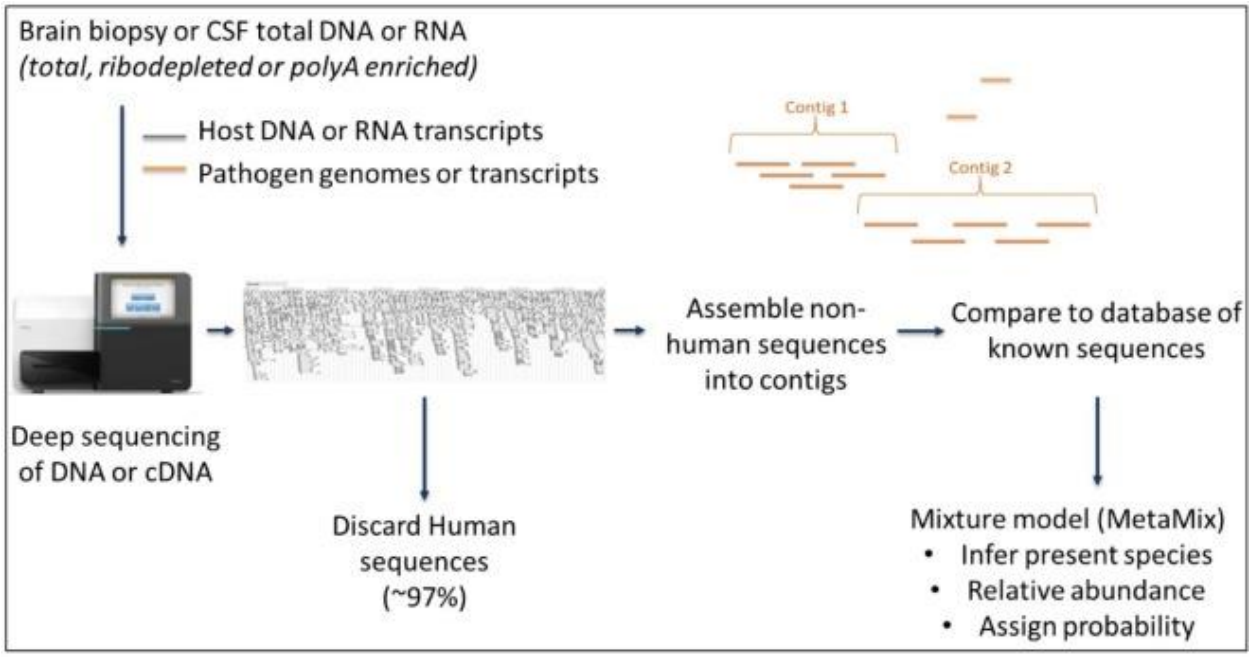
A method which has recently been applied to pathogen detection in cases of encephalitis is metagenomic analysis using next generation sequencing (NGS).

NGS, also known as deep sequencing, generates a single sequence from each fragment of DNA, or cDNA, present in a specimen. Downstream analysis allows differentiation between the origin of sequence fragments, for instance human, a specific bacterial species or a particular virus. This means mixed specimens, that contain host and microbial sequences, can be resolved.



Once sequences are generated, complex downstream bioinformatic analysis is required to identify the presence of any pathogen sequences.

In brief, any reads mapping to the human genome are removed, after which all remaining non-human sequences are compared to a database of known sequences to identify the provenance of the unknown sequences.



Results

Twenty-five articles were identified from the search. All the included articles were case reports, or case series of 1–7 patients. Altogether 44 cases were reported in which NGS provided a diagnosis.

Of the 22 cases that reported immune status of the patient, 73% (16/22) were immunocompromised. There was uniformly poor reporting of encephalitis or meningoencephalitis case definitions, and limited explanation of diagnostic assays performed and algorithms used for testing.

In 16 of the 44 known cases, causes of encephalitis were detected by rapid and specific primary screening methods such as PCR. Organisms included HSV, coxsackievirus A9, measles virus, VZV, mumps virus, Epstein-Barr virus, JC virus and Mycobacterium tuberculosis.

In the remaining 28 cases novel (18/44), rare (5/44) or unexpected (5/44) organisms were detected which could not been detected using specific PCR assays.

The five cases in which rare causes of encephalitis were identified were Brucella melitensis, Candida tropicalis, Leptospira santarosai and two cases of Balamuthia mandrillaris.

Advantage of using NGS for the diagnosis of encephalitis is that, aside from pathogen identification, in instances where virus titre and read depth is high enough it is possible to generate partial or full genome sequences for the pathogen.

Table 1. Reports of infectious encephalitis diagnosed by metage sequencing (mNGS) meeting inclusion criteria.

Case No.	Age (yrs)	Immunosuppressed?	Specimen Type	Pathogen Identified	Confirmatory testing of metagenomics result	Final Diagnosis	Type of pathogen	Treatment	Outcome
1	63	Yes (post solid organ transplantation)	Pooled RNA from brain, cerebrospinal fluid, serum, kidney, and liver	Arenavirus	Viral culture, EM, immunohistochemistry and serology. Donor IgM and IgG positive	Arenavirus Encephalitis	Novel organism	None	Died
2	64	Yes (post solid organ transplantation)	Pooled RNA from brain, cerebrospinal fluid, serum, kidney, and liver	Arenavirus	Viral culture, EM, immunohistochemistry and serology. Seroconversion	Arenavirus Encephalitis	Novel organism	None	Died
3	44	Yes (post solid organ transplantation)	Pooled RNA from brain, cerebrospinal fluid, serum,	Arenavirus	Viral culture, EM, immunohistochemistry and serology.	Arenavirus Encephalitis	Novel organism	None	Died

Temporal trends in the Publication of Encephalitis Cases involving Next-Generation Sequencing in the last Decade.