A Survey of Genome Sequencing Methods and Applications

Heba Abd-AlQader Abd-AlKareem Ibrahim

Under Supervision of: **Prof. Mohammed Abo Rizka**

College of Computing and Information Technology

الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحري

Arab Academy for Science, Technology & Maritime Transport

Abstract

Genome sequencing nowadays is a first step for understanding the genome. It can tell the scientists lots about what the genes are. It plays an important role in studying diseases, genetic disorders, and developing new treatments. This paper is to discuss the latest methodologies for sequencing genome, advantages, disadvantages, challenges and suggested enhancements.

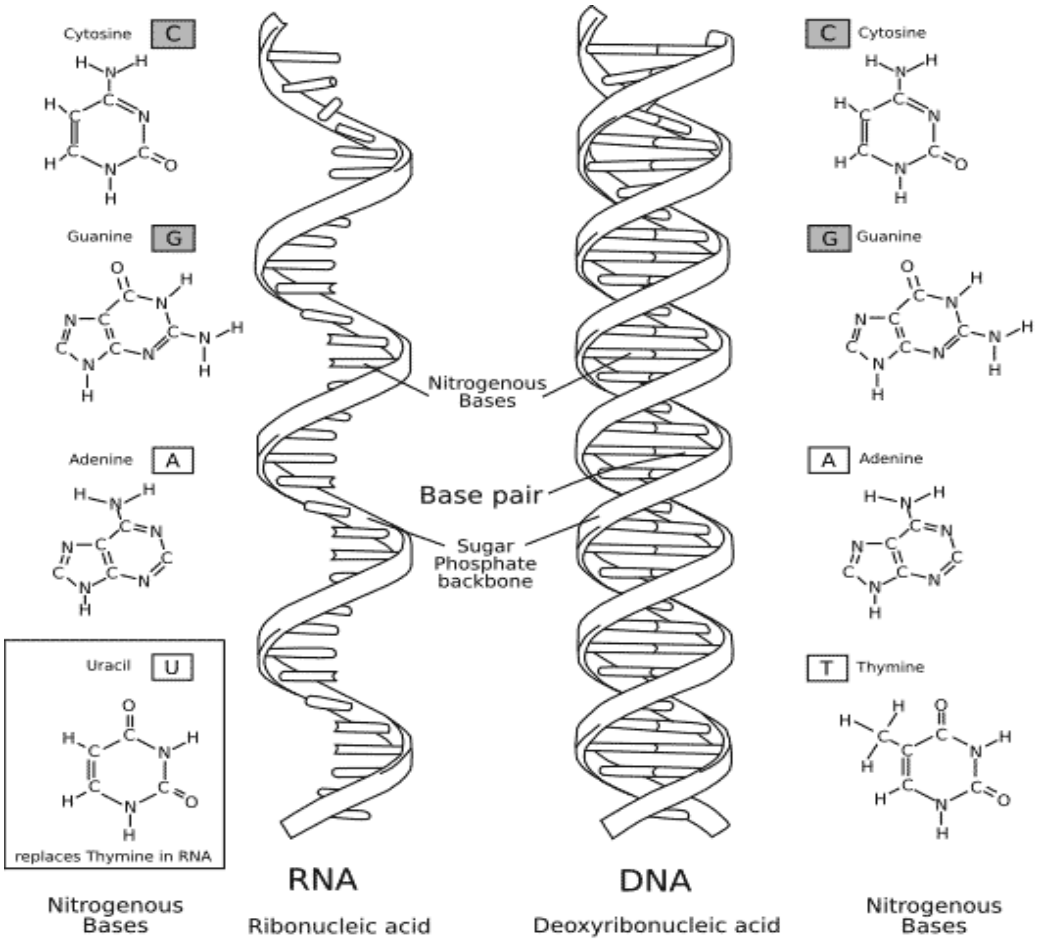
Keywords: Genome, DNA, Genome Sequencing, RNA

Background

A genome is the genetic material of a living organism. A genome is a complete set of DNA. It consists of genes which are packaged into chromosomes. They are holding all specific characteristics of an organism.

DNA is a molecule which is made of nucleotides which are strung together in a row. These nucleotides are four types which we call: Adenine, Thymine, Cytosine and Guanine abbreviated as A, T, C and G.

RNA is another molecule defining forms of like. It consists of nucleotides like DNA. Main difference between RNA and DNa is that DNA contains sugar deoxyribose while RNA contains different sugar ribose. Another difference is that RNA replaces Thymine with Uracil.



Introduction

A genome sequence is a long string of letters. Genome sequencing is getting the order of nucleotides in a genome of a living organism. It can look like this format:

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCG
GGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACC
GGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
AAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCT
CCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAG
GTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCG
CCGGGACAGAATGCCCTGAGGAACTTCTTCTGGAAGACCTTCTCCT
CCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATT
ACAGACCTGAA

Genome sequencing is very important to scientists as it is the first step for studying and understanding everything about the genome. It is the clue for understanding what genes are, their functionality, and what they say about the living organism.
Sequencing is considered a very challenging task as it produces massive amounts which requires time, processing, accuracy in detection and analysis. In this paper we are going to discuss the different technologies have been used in genome sequencing in the recent years discussing their advantages and disadvantages.

DNA sequencing history brief

[13] It all started back in 1953 when Watson and Crick discovered the DNA double helix structure. [15] They solved the three-dimensional structure of DNA, working from crystallographic data produced by Rosalind Franklin and Maurice Wilkins. However, the ability to 'read' or sequence DNA did not follow for some time. Strategies developed to infer the sequence of protein chains did not seem to readily apply to nucleic acid investigations: DNA molecules were much longer and made of fewer units that were more similar to one another, making it harder to distinguish between them. New tactics needed to be developed.

In 1965 the first whole nucleic acid was sequenced by Robert Holley and his colleagues, that of alanine tRNA from Saccharomyces cerevisiae. In parallel, Fred Sanger and his colleagues developed a related technique based on the detection of radio labelled partial-digestion fragments after two-dimensional fractionation, which allowed researchers to steadily add to the growing pool of ribosomal and transfer RNA sequences. It was also by using this 2-D fractionation method that Walter Fiers' laboratory was able to produce the first complete protein-coding gene sequence in 1972, that of the coat protein of bacteriophage MS2, followed four years later by its complete genome .

First-generation DNA sequencing

Chemical cleavage sequencing - by Maxam and Gilbert

The first DNA sequencing method appeared in 1977 was known as chemical cleavage sequencing, was published in February by Maxam and Gilbert. Chemical reactions are, indeed, the basis of this method. The method requires radioactive labelling at one end and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportions of one or two of the four nucleotide based in each of four reactions (G,A+G, C, C+T). Thus a series of labelled fragments is generated from the radio labelled end to the first 'cut' site in each molecule. The fragments in the four reactions are arranged side by side in gel electrophoresis for size separation. To visualize the fragments the gel is exposed to X-ray film for auto radiography yielding a series of dark bands each corresponding to a radio labelled DNA fragment from which the sequence may be inferred.

Sanger sequencing

In the Sanger method, the DNA strand to be analyzed is used as a template and DNA polymerase is used, in a PCR reaction, to generate complimentary strands using primers. Four different PCR reaction mixtures are prepared, each containing a certain percentage of dideoxynucleoside triphosphate (ddNTP) analogs to one of the four nucleotides (ATP, CTP, GTP or TTP). Synthesis of the new DNA strand continues until one of these analogs is incorporated, at which time the strand is prematurely truncated. Each PCR reaction will end up containing a mixture of different lengths of DNA strands, all ending with the nucleotide that was dideoxy labeled for that reaction.

Gel electrophoresis is then used to separate the strands of the four reactions, in four separate lanes, and determine the sequence of the original template based on what lengths of strands end with what nucleotide.

In the automated Sanger reaction, primers are used that are labeled with four different colored fluorescent tags. PCR reactions, in the presence of the different dideoxynucleotides, are performed as described above. However, next, the four reaction mixtures are then combined and applied to a single lane of a gel. The color of each fragment is detected using a laser beam and the information is collected by a computer which generates chromatograms showing peaks for each color, from which the template DNA sequence can be determined.

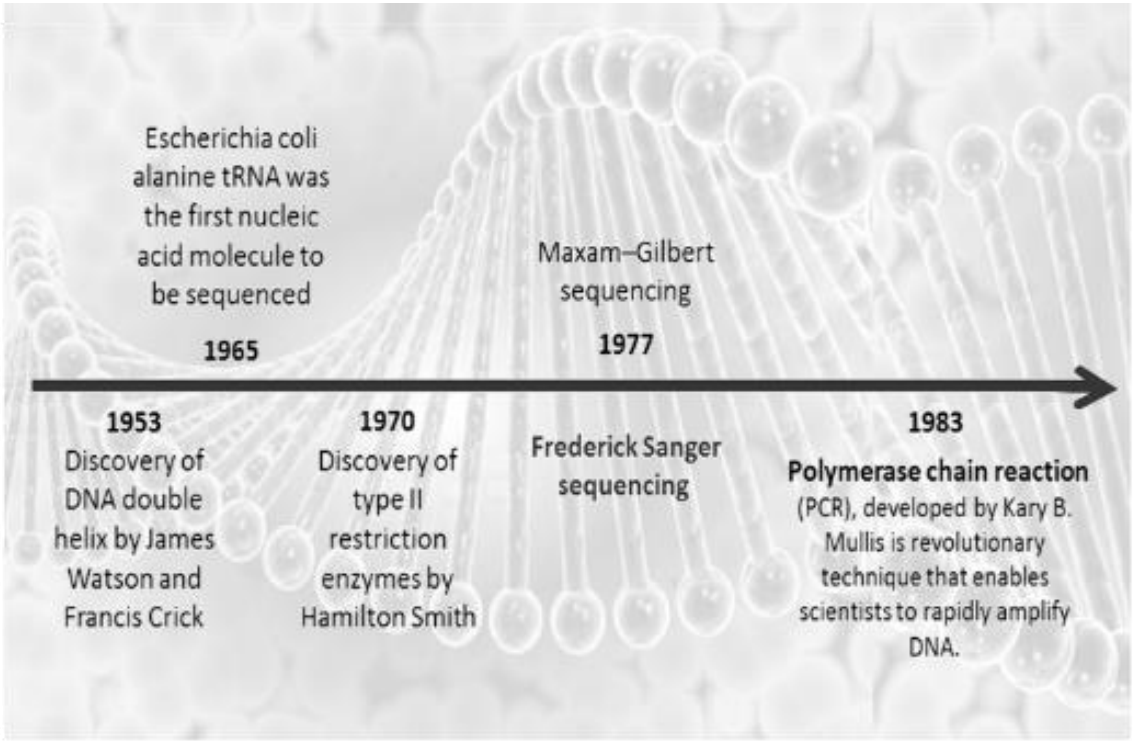
Typically, the automated sequencing method is only accurate for sequences up to a maximum of about 700-800 base-pairs in length. However, it is possible to obtain full sequences of larger genes and, in fact, whole genomes, using step-wise methods such as Primer Walking and Shotgun sequencing.

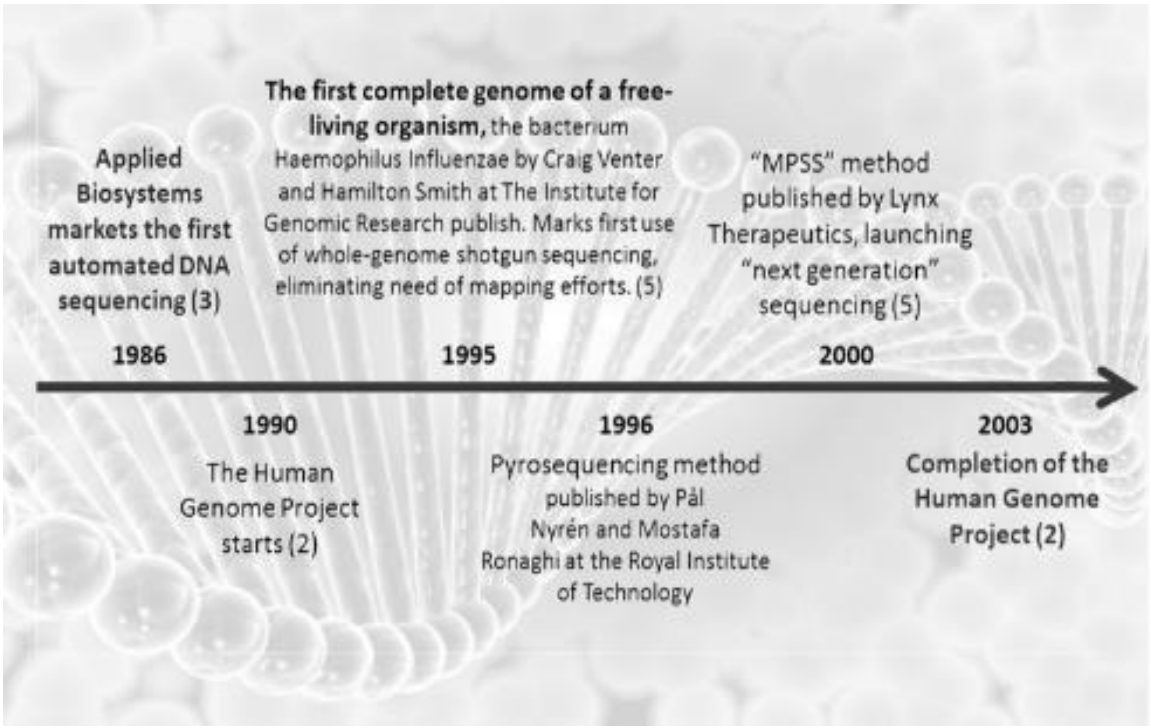
In Primer Walking, a workable portion of a larger gene is sequenced using the Sanger method. New primers are generated from a reliable segment of the sequence and used to continue sequencing the portion of the gene that was out of range of the original reactions.

Shotgun sequencing entails randomly cutting the DNA segment of interest into more appropriate (manageable) sized fragments, sequencing each fragment, and arranging the pieces based on overlapping sequences. This technique has been made easier by the application of computer software for arranging the overlapping pieces.

Sanger sequencing gives high-quality sequence for relatively long stretches of DNA (up to about 900 base pairs). It's typically used to sequence individual pieces of DNA, such as bacterial plasmids or DNA copied in PCR.

However, Sanger sequencing is expensive and inefficient for larger-scale projects, such as the sequencing of an entire genome or metagenome (the “collective genome” of a microbial community). For tasks such as these, new, large-scale sequencing techniques are faster and less expensive.


Escherichia coli alanine tRNA was the first nucleic acid molecule to be sequenced
1965
Maxam–Gilbert sequencing
1977
1953
Discovery of DNA double helix by James Watson and Francis Crick
1970
Discovery of type II restriction enzymes by Hamilton Smith
Frederick Sanger sequencing
1983
Polymerase chain reaction (PCR), developed by Kary B. Mullis is revolutionary technique that enables scientists to rapidly amplify DNA.


Applied Biosystems markets the first automated DNA sequencing (3)
1986
The first complete genome of a free-living organism, the bacterium Haemophilus Influenzae by Craig Venter and Hamilton Smith at The Institute for Genomic Research publish. Marks first use of whole-genome shotgun sequencing, eliminating need of mapping efforts. (5)
1995
"MPSS" method published by Lynx Therapeutics, launching "next generation" sequencing (5)
2000
1990
The Human Genome Project starts (2)
1996
Pyrosequencing method published by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology
2003
Completion of the Human Genome Project (2)

Second-generation DNA sequencing (Beginning of NGS - Next Generation Sequencing)

The key feature for these methodologies is parallelization of high number of sequencing reactions. This was achieved by miniaturization of sequencing reactions. In addition, detection systems were greatly improved. The time needed to determine the Gbp-sized sequences was reduced to hours or days, with the accompanying price reduction, compared to first generation sequencing methodologies.

Pyrosequencing

The pyrosequencing technique, pioneered by Pål Nyrén and colleagues, possessed a number of features that were considered beneficial: it could be performed using natural nucleotides (instead of the heavily-modified dNTPs used in the chain-termination protocols), and observed in real time (instead of requiring lengthy electrophoreses). It is based on the detection of light signal upon incorporation of nucleotide by polymerase. A parallelized version of pyrosequencing was developed by 454 Life Sciences, which has since been acquired by Roche Diagnostics. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs.This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other.

Roche 454

The sequencing machines produced by 454 Life Sciences (later purchased by Roche) were a paradigm shift in that they allowed the mass parallelization of sequencing reactions,

greatly increasing the amount of DNA that can be sequenced in any one run. It was the first NGS platform available as a standalone system.

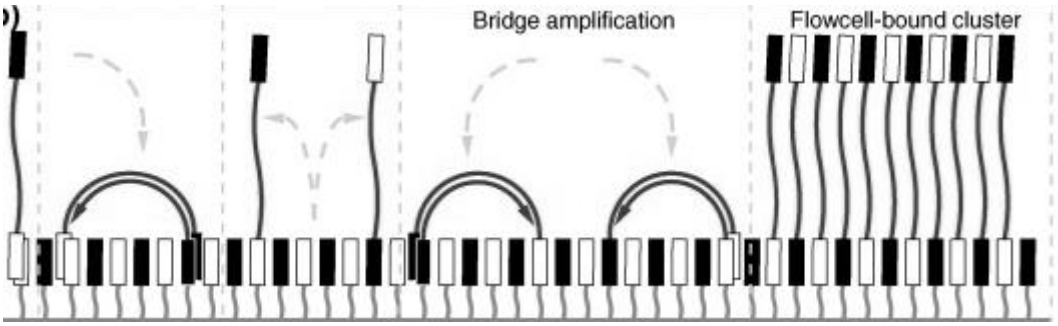
Libraries of DNA molecules are first attached to beads via adapter sequences, which then undergo a water-in-oil emulsion PCR (emPCR) to coat each bead in a clonal DNA population, where ideally on average one DNA molecule ends up on one bead, which amplifies in its own droplet in the emulsion. These DNA-coated beads are then washed over a picoliter reaction plate that fits one bead per well; pyrosequencing then occurs as smaller bead-linked enzymes and dNTPs are washed over the plate, and pyrophosphate release is measured using a charged couple device (CCD) sensor beneath the wells. This set up was capable of producing reads around 400–500 base pairs (bp) long, for the million or so wells that would be expected to contain suitably clonally-coated beads. This parallelization increased the yield of sequencing efforts by orders of magnitudes, for instance allowing researchers to completely sequence a single human's genome – that belonging to DNA structure pioneer, James Watson – far quicker and cheaper than a similar effort by DNA-sequencing entrepreneur Craig Venter's team using Sanger sequencing the preceding year.

Illumina sequencing

A number of parallel sequencing techniques sprung up following the success of 454. The most important among them is arguably the Solexa method of sequencing, which was later acquired by Illumina. This technique is able to produce larger volume of data, compared to 454 sequencer. With the read length of 100 bp, and run time of 11 days, the current instrument can produce 600 Gbp of data. It has become the most widely used NGS system in Plant biotechnology and breeding.

Illumina captures template DNA that has been ligated to specific adapters in a flow cell, a glass enclosure similar in size to a microscope slide, with a dense lawn of primers. The template is then amplified into clusters of identical molecules, or polonies, and sequenced in cycles using DNA polymerase. Terminator dNTPs in the reaction are labeled with different fluorescent labels and detection is by optical fluorescence. As only terminators are used, only one base can be incorporated in one cluster in every cycle. After the reaction is imaged in four different fluorescence levels, the dye and terminator group is cleaved off and another round of dye-labeled terminators is added. The total number of cycles determine the length of the read and is currently up to 101 or 151, for a total of 101 or 151 bases incorporated,

respectively. At the time of writing this review, this technology was able to yield the highest throughput of any system, with one of the highest raw accuracies.



One major disadvantage is the short read it produces. However, paired-end protocols virtually double the read per template and facilitate some applications that were originally out of the reach of the technology.
The Illumina HiSeq 2000 sequencer is currently able to sequence up to 540-600 Gbp in a single 2-flow cell, 8.5-day run at a cost of about 2 cents per Mbp.

SOLiD sequencing

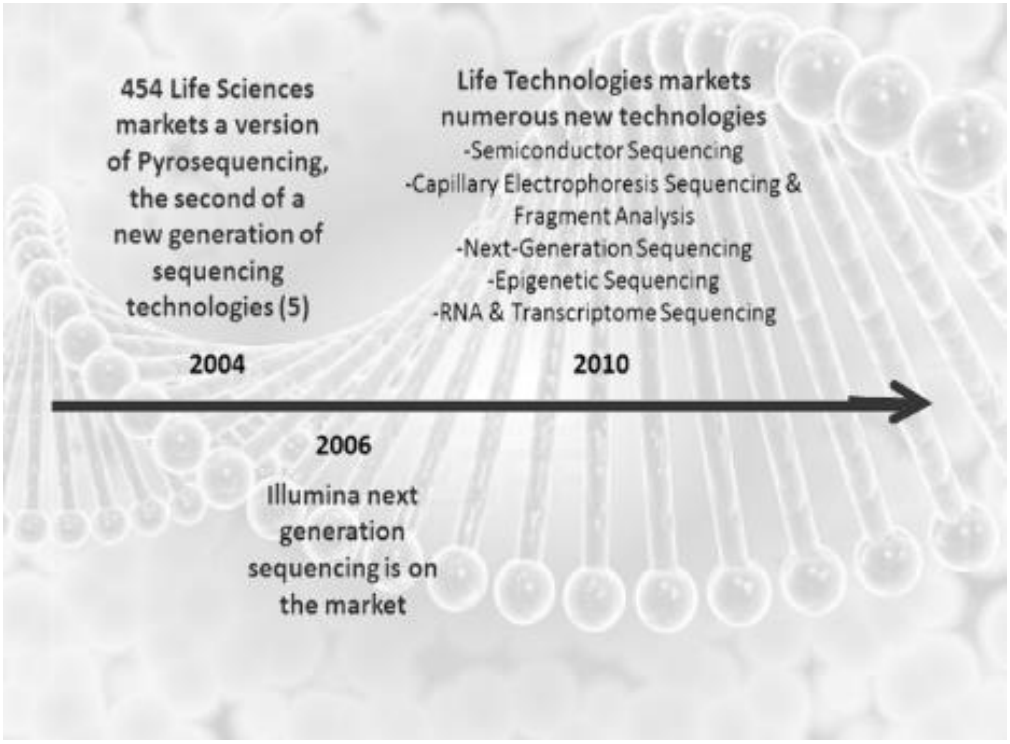
ABI (now part of Life Technologies) has commercialized the SOLiD (Support Oligonucleotide Ligation Detection) platform. This platform is based on Sequencing by Ligation (SbL) chemistry. SbL is a cyclic method but differs fundamentally from other cyclic NGS chemistries in its use of DNA ligase instead of polymerase, and two-baseencoded probes instead of individual bases as units. In this, a pool of all possible oligonucleotides of fixed length are labelled according to the sequenced position. This sequencing results to the sequences of quantities and lengths comparable to illumine sequencing.
While the SOLiD platform is not able to produce the read length and depth of Illumina machines, making assembly more challenging, it has remained competitive on a cost per base basis.

Third generation (emerging technologies)

NGS technologies provide high throughput at low cost compared to first generation sequencing approaches in much shorter time. Using these technologies an entire human genome can be sequenced within a single day. Data Sets produced by NGS are increasing in size are really a computational challenge. A single run of NGS machine produces terabytes of data.

NGS limitations are like the required infrastructure such as computer capacity and storage. Also the personnel expertise required to analyze the data is a challenge. In addition, the data needs to be managed wisely and skillfully as the resulted data is huge in amount and should be stored properly.

Third generation technologies are driven by the goal to reduce the price of sequencing and to simplify the procedure. One way to achieve this is to avoid preparation and amplification of samples by sequencing single molecules, thereby reducing the cost of reagents. Furthermore, optical systems for detection of incorporation events have inherent drawbacks, and another tendency of third generation methods is to use non-optical detection systems.

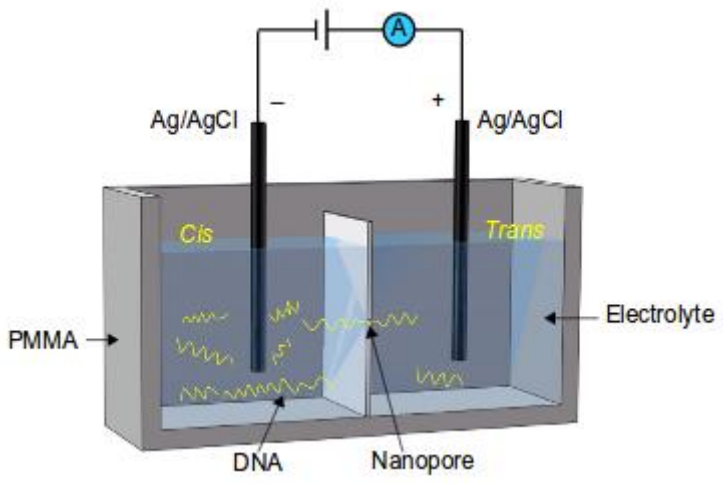


Single molecule real time (SMRT) sequencing

SMRT sequencing is based on the sequencing by synthesis approach. The DNA is synthesisd in so calles zero-mode wave-guides (ZMWs) - small well-like containers with the capturing tools located at the bottom of the well. The sequencing is performed with use of unmodified polymerase and fluorescently labelled nucleotides flowing freely in the solution. The wells are constructed in a way that only the fluorescence occurring by the bottom of the well is detected. The fluorescent label is detached from the nucleotide at its incorporation into the DNA strand, leaving an unmodified DNA strand. The SMTR technology allows detection of nucleotide modifications. This happens through the observation of polymerase kinetics. This approach allows reads of 1000 nucleotides. Reported accuracy is 99.3%, and it most likely could be improved by circularizing the template and sequencing it several times. Read length is more than 900 bp.

Nanopore DNA sequencing

Nanopore is a single-molecule sequencing technology which appeared for the first time in 1996 [7] which offers low-cost genotyping. This technology has many application in the analysis of RNA, DNA, ions, proteins, drugs and others. Nanopores are divided into two types: Biological Nanopores and Solid State Nanopores. Also there are the Hybrid biological/solid-state nanopores. Both categories can be used for sensing the biological and chemical molecules.



Schematic view of nanopore fluidic setup comprising a DNA molecule through the nanopore (PMMA stands for polymethyl methacrylate)

Nanopore DNA sequencing is considered one of the most powerful technologies. The advantages for this technologies are significant as it includes that it is label-free, high throughput, low material requirement and very long reads. It is a low cost and fast DNA sequencing technology.

A disadvantage of Nanopore DNA sequencing technology is the very high error rate which is over 90%. The main challenge for this technology is refining the resolution of the detected bases.

Ion Torrent

Ion Torrent (another Life Technologies product) is the first so-called 'post-light sequencing' technology, as it uses neither fluorescence nor luminescence. It was developed by the inventor of 454 sequencing. Sample preparation is very similar to the one used for 454 sequencing. How - ever, instead of PicoTiterPlate, a chip with ion-sensitive field-effect transistor sensor, engineered to be able to detect individual protons, is used. The instrument has no optical components. Beads containing enriched template DNA are deposited into the wells of the chip. The chip is situated within the flow cell, and is sequentially flushed with individual dNTPs. Integration of nucleotide releases H+ that changes the pH of the surrounding solution proportional to the number of incorporated nucleotides. This is detected by the sensor on the bottom of each well, and converted to electronic signal, that is re - corded by the system. The latest version of the instrument (Ion Proton) is claimed to have throughput of 20 Gbp, with reads of 200 bp, and run time of up to 4 h.

The major disadvantages of the system are problems in reading homopolymer stretches and relative short read length, what is partially compensated by the number of wells (up to 11 million).

The major advantage seems to be the price.

Sequencing Platform Comparison

Platform	Amplification	Sequencing	Detection	Read Length	Output per Run	Run Time
Second-generation sequencing platforms						
454	Emulsion PCR on beads	Unlabeled nucleotide incorporation	Detection of light emitted by release of PP_i	Variable (400 bp for single end sequencing)	400--600 Mbp	10 h
SOLiD	Emulsion PCR on beads	Ligation of 2-base encoded fluorescent oligonucleotides	Fluorescence emission from labeled oligonucleotides	75+35 bp	20--30 Gbp	7 d
Illumina	Array-based enzymatic amplification	Fluorescently labeled end-blocked nucleotide incorporation	Fluorescence emission from nucleotides	2[times]100 bp	100--200 Gbp	8 d
Third-generation sequencing platforms						
Oxford nanopore	NA	Processive endo- or exonuclease activity feeds individual bases or whole DNA	Current disruption across nanopore corresponds to	Variable	Variable	Variable

		strands through protein or solid-state nanopores	nucleotide structure			
Ion Torrent	Variable	DNA polymerase incorporation of unlabeled nucleotides added sequentially to solid-state microwells	Solid-state detection of hydrogen ions released by nucleotide incorporation	200 bp	10 Mbp to 1 Gbp	2 h

bp indicates base pair; Gbp, one billion base pairs; Mbp, one million base pairs; PP_i, pyrophosphate; and PCR, polymerase chain reaction.

NGS technologies provide opportunities for understanding unknown species and complex diseases. Although different companies implement different platforms with distinctive features and advantages, depend on the number of reads and the read length to ensure assembly quality and accuracy. Therefore, an important issue for future research will be the improvement of methods used for analysis of the huge amount of data produced by NGS. The goals will be to increase the accuracy of assembly sequencing, reduce the processing time, and fine-tune the efficiency of algorithms for analysis. In order to make the best use of NGS data, the design of state-of-the-art bioinformatics pipelines to extract meaningful biological insights will be a significant topic in the following years. Ultimately, NGS could reveal human genomic information and help to elucidate the function of the genome, which may provide therapeutic regimens for personalized medicine in the future.

Genome sequencing applications

Due to its power to provide comprehensive information on the genetic sequence of organisms, sequencing has been used for many amazing applications and experiments.

Decoding complexities of human disease

DNA Sequencing technology can help decode the complexities of human disease. When the Ebola outbreak threatened West Africa, Next Generation Sequencing technology enabled researchers to track the evolution of the Ebola virus in real time. Cancer, which is a multitude of different diseases with many characteristics, is the focus of a recent effort from the National Cancer Institute (NCI). Up to 1000 patients will have their tumors sequenced for genetic mutations that can be exploited by targeted therapeutics. Thus, patients could benefit from personalized genomic treatment options, tailored to their tumors.

Uncovering the secrets of our past

DNA derived from human specimens thousands of years old can now be sequenced to unveil insights into our evolutionary past. One recent study gives insights into how the Americas may have become populated. Over 16,000 years ago, a land bridge connected Siberia to Alaska, allowing migration of the first people from Asia to the Americas. There are many theories as to how people populated the North and South of America. However, recent next generation sequencing data has suggested an interesting model. In an ancient burial site in Alaska, two preserved infants were found and their mitochondrial genome (the genetic sequence of the energy-producing organelle found in every cell in the body and is maternally inherited) was sequenced. Interestingly, the mitochondrial DNA from the two infants was very different from that of modern people living in North America—and was more closely similar to mitochondrial DNA from a 500 year-old Incan child mummy found in Argentina. This has huge implications as to how humans populated the Americas, suggesting that the people who first arrived in Alaska from Asia were already very diverse and that they traversed all the way down the American continent to South America.

Validating herbal supplements

Many people are turning to herbal supplements, seeking a more natural way to stay healthy. This has led to the dietary supplement market ballooning to an estimated worth of over $123 billion dollars worldwide. Although widely consumed, data from next generation sequencing has questioned whether some herbal supplements actually contained the ingredients on the label. Researchers at the University of Guelph in Ontario, Canada used DNA barcoding to sequence the contents of 44 herbal supplements. Their results were horrifying: a third of the samples did not even contain the advertised plant, with many containing ingredients not listed on the bottle. Some even contained other supplements that were instead powerful laxatives or were mixed with substances that can induce an allergic reaction in many people, such as nuts. Only two of the products tested contained 100% authentic ingredients.

Conserving our wildlife

Sequencing is also aiding conservation efforts around the world. For example, Thermo Fisher Scientific recently donated a Genetic Analyzer to the Cheetah Conservation Fund (CCF), Life Technologies Conservation Genetics Laboratory: the only fully equipped genetics lab at a conservation facility in Africa. Due to human-wildlife conflict, the cheetah population has declined by 90% in the 20th century alone, and the International Union for the Conservation of Nature (IUCN) lists them as "vulnerable." Sequencing enables the researchers at CCF to better understand the underlying genetics of this beautiful big cat and monitor the reproduction and health of the cheetah population. This information is crucial to ensure that we do not lose the Cheetah forever.

In agriculture

DNA sequencing plays vital role in the field of agriculture. The mapping and sequencing of the whole genome of microorganisms has allowed the agriculturists to make them useful for the crops and food plants. For example, specific genes of bacteria have been used in some food plants to increase their resistance against insects and pests, as a result the productivity and nutritional value of the plants may increase.

In forensic science

DNA sequencing is used to identify the criminals by finding some proof from the criminal scene in the form of hair, nail, skin, or blood samples. DNA sequencing can also be used to determine the paternity of the child.

In medical research, DNA sequencing can be used to detect the genes which are associated with some hereditary or acquired diseases.

DNA sequencing information is important for planning the procedure and method of gene manipulation.

DNA sequencing is used to construct the molecular evolution map.

Bioinformatics issues and challenges

Data storage and analysis

As sequencing technologies advance, biologists are slowly drowning in their data. A typical second generation sequencing run produces some terabytes of image data, from which some gigabases of sequence is extracted. Once the data comes off the sequencer and has been saved to a database, the challenge begins: How to analyze the data in a fast, accurate and meaningful way? Much work is currently focused on fast algorithms for mapping reads to reference sequences and for assembling reads.

The BAM file (a semicompressed alignment file) for a single 30X human whole-genome sample is about 90 GB. A relatively modest project of 100 samples would generate 9 TB of BAM files.

With a single Illumina HiSeq X instrument capable of generating over 130 TB of data per year, storage can quickly become a concern. For example, the Broad Institute is generating sequencing data at the rate of one 30X genome every 12 minutes—nearly 4,000 TB worth of BAM files every year.

Identifying Redundant Sequences

It is very important to identify redundant sequences – duplicate reads, which are a result of PCR amplification. These duplicates must be removed before variant calling! Remember also that PCR amplification may introduce sequencing errors – if your PCR introduces one little nucleotide change, and then amplifies it, you may end up with a decrease variant detection and sensitivity, as you will be introducing something new to your sequence, and have it massively represented. It is very difficult to recognize sequences that were altered by PCR. Therefore they represent a big risk in your experiment (especially if you are looking for something that is not well represented in your sample). A safe way to recognize it is to remove duplicates, and/or run a Sanger sequencing with the same sample (always remember to check your PCR primers to make sure there isn't any chance of allele dropout – but that's another topic altogether!). If you are confident in your Sanger sequencing, it doesn't show up

in your electropherogram, and after you remove the duplicates it only appears on one sequence, you can be more confident that is was only a PCR artifact.
So, every time you have the same sequence, start site and orientation you may have multiple reads of the same unique DNA fragment, and you must remove them before continuing analyses.

Short Reads vs. Long Reads

Some sequencing applications, such as the detection of single nucleotide polymorphisms, can be managed with short-read technology. Other applications, such as the detection of structural variants, may demand long-read technology, and some applications, such as the assembly of a new organism's genome, may require a combined approach, with short reads providing accuracy and high throughput, where possible, and long reads coping with highly repetitive genomic regions.
Illumina's dominance of the sequencing market has meant that the vast majority of the data that has been generated so far is based on short reads. Having a large number of short reads is a good fit for a number of applications, such as detecting single-nucleotide polymorphisms in genomic DNA and counting RNA transcripts. However, short reads alone are insufficient in a number of applications, such as reading through highly repetitive regions of the genome and determining long-range structures.
Long-read platforms, such as the RSII and Sequel from Pacific Biosciences and the MinION from Oxford Nanopore Technologies, are routinely able to generate reads in the 15–20 kilobase (kb) range, with individual reads of over 100 kb having been reported. Such platforms have earned the respect of scientists such as Charles Gasser, Ph.D., professor of molecular and cellular biology at the University of California, Davis.
"I am impressed with the success people have had with using the long-read methods for de novo genome assembly, especially in hybrid assemblies when combined with short-read higher fidelity data," comments Dr. Gasser. "This combination of technologies makes it possible for a single investigator with a very small group and a minimal budget to produce a useable assembly from a new organism's genome."

To get the most out of these long-read platforms, however, it is necessary to use new methods for the preparation of DNA samples. Standard molecular biology methods haven't been optimized for isolating ultra-long DNA fragments, so special care must be taken when preparing long-read libraries.

For example, vendors have created special “high molecular weight” kits for the isolation of DNA fragments >100 kb, and targeted DNA protocols have been modified to selectively enrich for large fragments of DNA. These new methods and techniques need to be mastered to ensure maximum long-read yield.

As an alternative to true long reads, some are turning to a specialized form of short reads called linked-reads, such as those from 10X Genomics. Linked-reads are generated by adding a unique barcode to each short read generated from a single long DNA fragment, which is generally >100 kb. The unique barcodes are used to link together the individual short reads during the analysis process. This provides long-range genomic information, enabling the construction of large haplotype blocks and elucidation of complex structural information.

“Short-read sequencing, while immensely powerful because of high accuracy and throughput, can only access a fraction of genomic content,” advises Dr. Saxonov. “This is because genomes are substantially repetitive and much of the information in the genome is encoded at long scales.”

References

1. Genome News Network, ch1, ch2 : http://www.genomenewsnetwork.org/resources/whats_a_genome
2. DNA and RNA - Computational Medical Center, Thomas Jefferson University: cm.jefferson.edu/learn/dna-and-rna

3. Daniel C. Koboldt Li Ding Elaine R. Mardis Richard K. Wilson (2010): Challenges of sequencing human genomes - academic.oup.com/bib/article/11/5/484/264316
4. Simone Roeh, Peter Weber, Monika Rex-Haffner, Jan M. Deussing, Elisabeth, B. Binder & Mira Jakovcevski (2017): Sequencing on the SOLiD 5500xl System– in-depth characterization of the GC bias - www.tandfonline.com/doi/abs/10.1080/19491034.2017.1320461#aHR0cHM6Ly93d3cudGFuZGZvbmxpbmUuY29tL2RvaS9wZGYvMTAuMTA4MC8xOTQ5MTAzNC4yMDE3LjEzMjA0NjE/bmVlZEFjY2Vzcz10cnVlQEBAMA==
5. Chial, H. (2008) DNA sequencing technologies key to the Human Genome Project - www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828
6. Artem Tarasov Albert J. Vilella Edwin Cuppen Isaac J. Nijman Pjotr Prins (2015): Sambamba: fast processing of NGS alignment formats - academic.oup.com/bioinformatics/article/31/12/2032/214758
7. Yanxiao Feng, Yuechuan Zhang, Cuifeng Ying, Deqiang Wang, Chunlei Du 'Nanopore-based Fourth-generation DNA Sequencing Technology' [GPB 144 (2015) – GPB 13/1 (4–16)] Genomics, Proteomics & Bioinformatics, Volume 13, Issue 6, December 2015, Pages 383: www.sciencedirect.com/science/article/pii/S1672022915000133
8. Ian C. Nova, Ian M. Derrington, Jonathan M. Craig, Matthew T. Noakes, Benjamin I. Tickman, Kenji Doering, Hugh Higinbotham, Andrew H. Laszlo, Jens H. Gundlach (2017): Investigating asymmetric salt profiles for nanopore DNA sequencing with biological porin MspA - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181599
9. https://www.healio.com/hematology-oncology/learn-genomics/whole-genome-sequencing/whole-genome-sequencing-methods
10. https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing
11. https://www.thebalance.com/dna-sequencing-methods-375671
12. DNA SEQUENCING – METHODS AND APPLICATIONS - http://library.umac.mo/ebooks/b28050393.pdf
13. James M.Heather, BenjaminChain (January 2016): The sequence of sequencers: The history of sequencing DNA - https://www.sciencedirect.com/science/article/pii/S0888754315300410
14. History of DNA sequencing technologies - http://dnasequencing.yolasite.com/timeline-and-history.php
15. THE HISTORY OF DNA SEQUENCING - https://www.dmbj.org.rs/jmb/pdf/2013-4/3.pdf
16. Maxam-Gilbert Sequencing: What Was It, and Why It Isn't Anymore - https://bitesizebio.com/36696/maxam-gilbert-sequencing/
17. DNA sequencing: Clinical applications of new DNA sequencing technologies - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3364518/
18. Decoding the Genome: Applications of DNA Sequencing - https://bitesizebio.com/27892/decoding-the-genome-applications-of-dna-sequencing/

19. DNA Sequencing: Methods and Applications - https://link.springer.com/chapter/10.1007/978-81-322-1554-7_2
20. https://ab.inf.uni-tuebingen.de/teaching/ss09/gbi/script/13-sequencing.pdf