Journal of Entomology and Zoology Studies

Available online at www.entomoljournal.com

E-ISSN: 2320-7078
P-ISSN: 2349-6800
JEZS 2018; 6(3): 1601-1609
© 2018 JEZS
Received: 18-03-2018
Accepted: 23-04-2018

Vandana Yadav
Ph.D. Scholar, Animal Genetics, ICAR-NDRI, ICAR-National Dairy Research Institute, Karnal, Haryana, India

BL Saini
Ph.D Scholar, Animal Genetics, ICAR-IVRI, ICAR-National Dairy Research Institute, Karnal, Haryana, India

Narendra Pratap Singh
M.V.Sc Scholar, Animal genetics, ICAR- NDRI, ICAR-National Dairy Research Institute, Karnal, Haryana, India

Neeti Lakhani
Ph.D Scholar, Nutrition division, ICAR-NDRI, ICAR-National Dairy Research Institute, Karnal, Haryana, India

Rahul Sharma
VEO, DAH, Katni, Madhya Pradesh, India

Correspondence
Vandana Yadav
Ph.D. Scholar, Animal Genetics, ICAR-NDRI, ICAR-National Dairy Research Institute, Karnal, Haryana, India

Next generation sequencing and its application in livestock

Vandana Yadav, BL Saini, Narendra Pratap Singh, Neeti Lakhani and Rahul Sharma

Abstract
The ability to determine nucleic acid sequences is one of the most important platforms for the detailed study of biological systems. Next generation sequencing (NGS) technology has revolutionized genomic and genetic research. Starting from the basic Sanger sequencing, the next generation sequencing are providing high throughput and much cheaper alternative. All NGS have a similar base methodology that includes template preparation, sequencing and imaging, and data analysis. Different NGS platform use different technology for the identification of the called base. Different commercially available technologies include 454 FLX Roche (pyrosequecing), Illumina Solexa (CRT), ABI SOLiD (sequencing by ligation) and Ion semiconductor sequencing (Ion Torrent) which require the pre-amplification process before sequencing. Single molecule approach are the developing technology for sequencing which can provide the @1000$ per genome. Single molecule technology includes Helicos sequencer, SMRT, Nanopore technology that does not require the amplification prior to sequencing. Data analysis can be time-consuming and may require special knowledge of bioinformatics to gather accurate information from sequence data. Resequencing of the human genome is being performed to identify genes and regulatory elements involved in pathological processes. NGS has also provided a wealth of knowledge for comparative biology studies through whole genome sequencing of a wide variety of organisms. Simply analyzing all DNA from control versus infected (diseased) individuals will reveal the "extra" DNA, which most likely derives from the infectious agent. The approach was used successfully to identify e.g. colony collapse disorder killing honeybees but also to identify the cause of diseases that killed thousands of humans in the past. Thus NGS can be used for genomics, trancriptomics, epigenetic study and metagenomics study.

Keywords: Sanger sequencing, nucleotide bases, template, adapter

1. Introduction
DNA sequencing is the process of determining the precise order of nucleotide bases (adenine, guanine, cytosine, and thymine) within a DNA molecule. The advancement in DNA sequencing technology has been proved to be a boon for biological and medical research and discovery.

Knowledge of DNA sequences is not only required for basic biological research but in numerous applied fields such as medical diagnosis, biotechnology, forensic biology, virology and biological systematic. The modern DNA sequencing technology with rapid speed has been instrumental in the sequencing of the complete DNA sequences, or genomes of numerous types and various life forms, including the human genome and other complete DNA sequences of many animal, plants, and microbial species.

2. History of sequencing
Ray Wu at Cornell University (1970) established the first method for determining DNA sequences, based on a location-specific primer extension strategy. The cohesive ends of lambda phage DNA were sequenced by using DNA polymerase catalysis and specific nucleotide labeling, both of which forms the base for the current sequencing technologies [1]. This primer-extension strategy was then used to develop more rapid DNA sequencing methods. Frederick Sanger (1977) at the MRC Centre, Cambridge, UK then developed the method for "DNA sequencing with chain-terminating inhibitors". In the same year Walter Gilbert and Allan Maxam at Harvard also developed another method for DNA sequencing known as "DNA sequencing by chemical degradation". In 1987, Applied Biosystems gave the first full-automated sequencer (ABI370) for the first time.

After that various high throughput-sequencing platforms (Pyrosequencing-2005, Illumina-2006, Ion torrent-2011, Pac-Bio- 2012, illumine HiseqX-2013, Nanopore- 2014) were launched, marking the era of sequencing. The advancement in sequencing technologies has now made it possible to sequence the complete genome within a short time and in a cost effective manner.

3. Next-generation sequencing methods

Generation refers to the chemistry and technology used by the sequencing process. First generation generally refers to Sanger sequencing. "Next-generation" is generally used to refer to any of the high-throughput methods which were developed after Sanger. Third-generation refers to single-molecule methods (Helicos and Pac-Bio SMRT) and fourth generation includes the Nano-pore sequencing (Fig: 1).

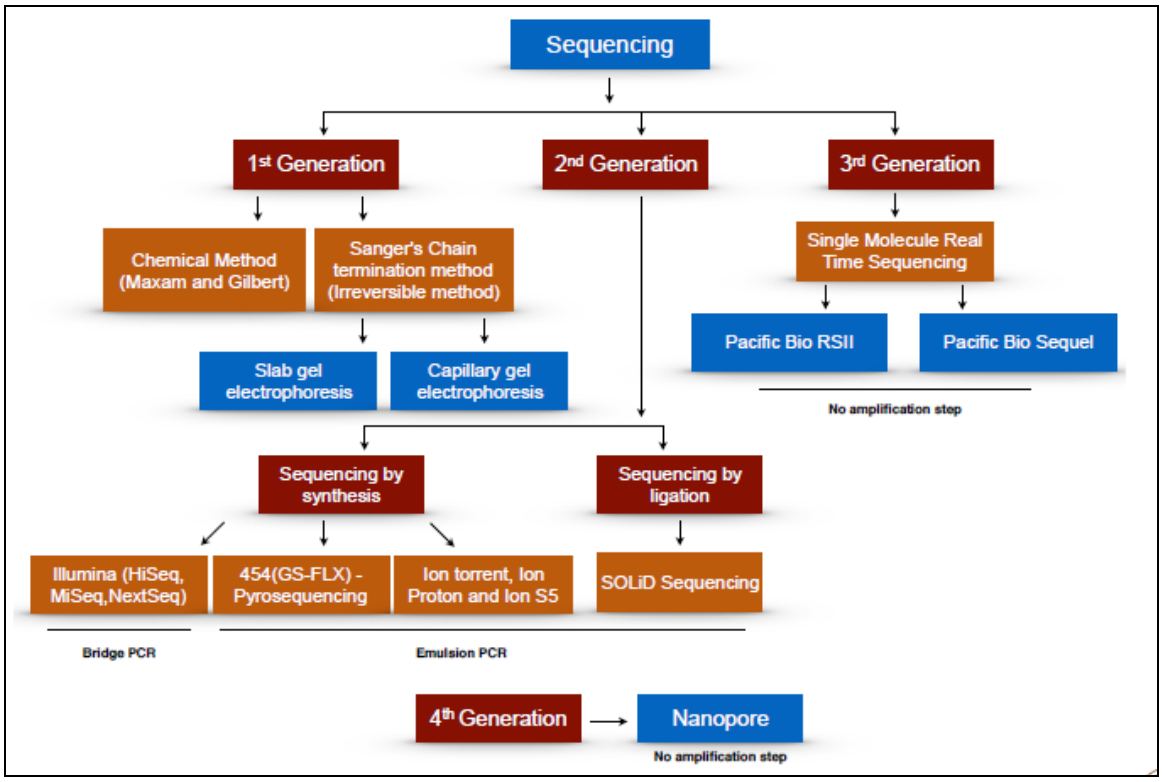

Fig 1: Different NGS platforms

3.1 Sanger method of sequencing

It was the most common approach used for DNA sequencing invented by Frederick Sanger in 1977 & got Nobel Prize in 1980.

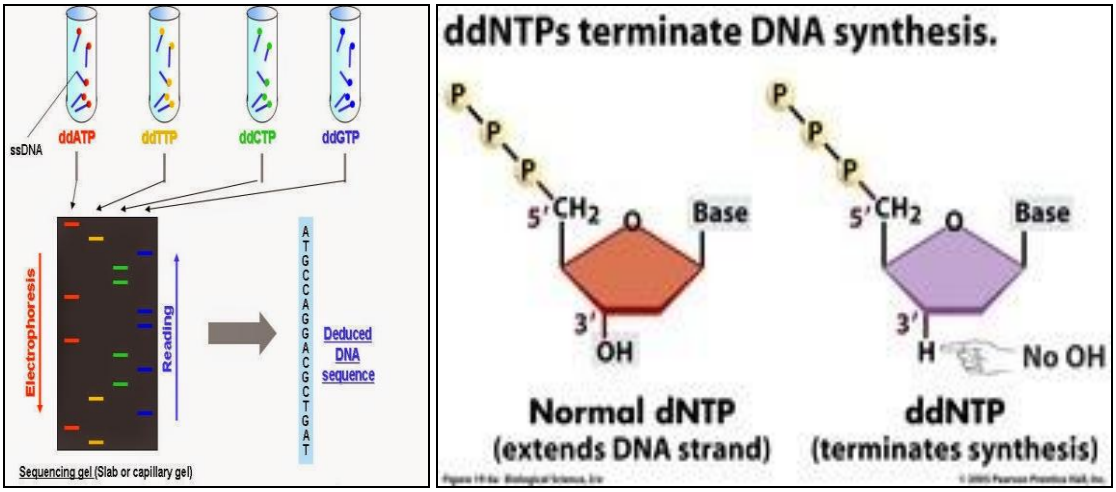

Fig 2: Sanger sequencing by chain termination

This method is also termed as Chain Termination or Dideoxy method (Fig: 2). It relies on the use of dideoxyribonucleoside triphosphate, derivative of the normal deoxyribonucleoside triphosphate that lack the 3′ hydroxyl group [2]. It is based on the synthesis of DNA in a controlled way to generate fragments terminating at specific point. Purified DNA is synthesized *in vitro* in a mixture that contains single stranded molecules of the DNA to be sequenced, the enzyme DNA polymerase, a short primer DNA to enable the polymerase to start DNA synthesis, and the four deoxyribonucleoside triphosphate (dATP, dCTP, dGTP, dTTP) [3]. If a dideoxyribonucleotide analog of one of these nucleotides is also present in the mixture, it can incorporate into a growing DNA chain. Because this chain now lacks a 3′ OH group, the addition of next nucleotide is blocked, and the DNA chain terminates at that point.

To determine the complete sequence of a DNA fragment, the double- stranded DNA is first separated into its single strands and one of the strands is used as a template for sequencing. Four different chain- terminating dideoxyribonucleoside triphosphates (ddATP, ddCTP, ddGTP, ddTTP) are used in four separate DNA synthesis reactions on copies of the same single-stranded DNA template. Each reaction produces a set of DNA copies that terminate at different points in the sequence. The products of these four reactions are separated by electrophoresis in four parallel lanes of a thin slab polyacrylamide gel. The newly synthesized fragments are detected by a label (either radioactive or fluorescent) that has been incorporated either into the primer or into of the deoxyribonucleoside triphosphates used to extend the DNA chain [4]. In each lane, the bands represent fragments that have been terminated at a given nucleotide but at different positions in the DNA. By reading off the bands in order, starting at the bottom of the gel and working across all lanes, the DNA sequence of the newly synthesized strands can be determined. Thus, a long and labor-intensive process was completed, and the sequencing data for the DNA of interest were in hand and ready for assembly, translation to amino acid sequence, or other types of analysis.

4. Advances in DNA sequencing technologies

The landmark publications of the late 1970 s by Sanger's and Gilbert's groups and notably the development of the chain termination method by Sanger and colleagues established the groundwork for decades of sequence-driven research that followed. The chain-termination method published in 1977 also commonly referred to as Sanger or dideoxy sequencing, has remained the most commonly used DNA sequencing technique to date and was used to complete human genome sequencing initiatives led by the International Human Genome Sequencing Consortium and Celera Genomics. Very recently, the Sanger method has been partially supplanted by several "next-generation" sequencing technologies that offer dramatic increase in cost-effective sequence throughput, albeit at the expense of read lengths [5]. The next-generation technologies commercially available today include the 454 GS20 pyrosequencing based instrument (Roche Applied Science), the Solexa 1G analyzer (Illumina, Inc.), the SOLiD instrument from Applied Biosystems, and the Heliscope from Helicos, Inc. and Nanopore from Oxford [6].

5. Workflow of next-generation sequence

The workflow to produce next-generation sequence-ready libraries is straightforward and includes sample preparation, library construction, clonal amplification, sequencing and data analysis (Fig: 3).

a. Sample preparation

- Source of sample (DNA, RNA)
- Qualify and quantify samples

b. Library Construction

- Prepare platform specific library.
- Fragmentation of target DNA and adapter ligation.

c. Clonal Amplification

- Bridge PCR amplification-Illumina sequencing.
- Emulsion PCR-SOLiD, Ion torrent and pyrosequencig.
- Third and fourth generation sequencing technologies don't require amplification.

d. Sequencing

- Perform sequencing run reaction on NGS platform.
- Sequencing by ligation-SOLiD sequencing.
- Sequencing by synthesis-Illumina sequencing, pyrosequencig, Ion torrent.

e. Data Analysis

- Application specific data analysis pipeline.

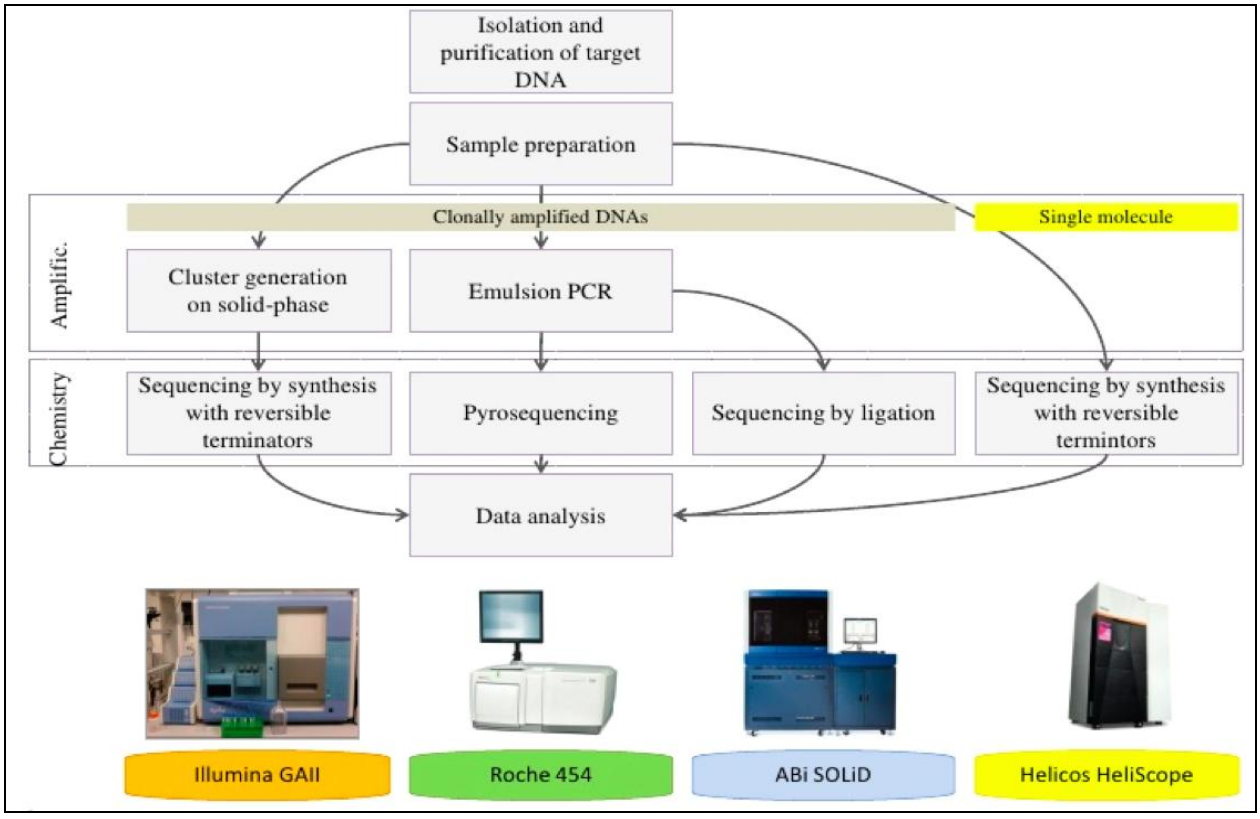

Fig 3: Workflow of NGS

DNA fragments that may originate from a variety of front-end processes are prepared for sequencing by ligating specific adaptor oligos to both ends of each DNA fragment. Importantly, relatively little input DNA (a few micrograms at most) is needed to produce a library. These platforms also have the ability to sequence the paired ends of a given fragment, using a slightly modified library process. This approach can be used if a de novo genome sequence is to be assembled from the next generation data. Finally, next-generation sequencers produce shorter read lengths (35-250 bp, depending on the platform) than capillary sequencers (650–800 bp), which also can impact the utility of the data for various applications such as de novo assembly and genome resequencing.

6. Pyrosequencing

An inherent limitation of Sanger sequencing is the requirement of *in vivo* amplification of DNA fragments that are to be sequenced, which is usually achieved by cloning into bacterial hosts. The cloning step is prone to host-related biases, is lengthy, and is quite labor intensive. The 454 technology, the first next-generation sequencing technology is based on a highly efficient *in-vitro* DNA amplification (emulsion PCR) rather than *in-vivo* amplification by cloning. The DNA fragment to be sequenced is fragmented and streptavidin beads are attached to each fragment using adapters. During amplification by emulsion PCR these fragments carrying streptavidin beads are captured into separate emulsion droplets, acting as individual amplification reactors and producing ~10^7 clonal copies of a unique DNA template per bead. The DNA templates with beads are then transferred to wells of a picotiter plate (one template per well) and analyzed using a pyrosequencing reaction.

Pyrosequencing, sequencing-by-synthesis method detecting successful incorporation of nucleotide as emitted photons. Since the single-stranded DNA fragments on the beads have been amplified with general tags, an universal primer is annealed permitting an elongation towards the bead. The emission of photons upon incorporation depends on a series of enzymatic steps. Incorporation of a nucleotide by a polymerase releases a diphosphate group (PPi), which catalyzed by ATP sulphurylase forms adenosine triphosphate (ATP) by the use of adenosine phosphosulphate (APS). Finally, the enzyme luciferase (together with D-luciferin and oxygen) can use the newly formed ATP to emit light. Another enzyme, apyrase, is used for degradation of unincorporated dNTPs as well as to stop the reaction by degrading ATP (Fig: 4).

In the 454 systems, the Pyrosequencing technology is adapted as follows. The enzymes luciferase and ATP sulphurylase are immobilized on smaller beads surrounding the larger amplicon carrying beads. All other reagents are supplied through a flow allowing reagents to diffuse to the templates in the Picotiter Plate. Polymerase and one exclusive dNTP per cycle generate one or more incorporation events and the emitted light is proportional to the number of incorporated nucleotides. Photons are detected by a CCD camera and after each round; apyrase is flowed through in order to degrade excess nucleotides. The washing procedure for the removal of byproducts permits read lengths of over 400 bp (250 bp in the GS FLX system and over 400 bp in the recently upgraded instrument, the GS FLX Titanium). This limitation is due to negative frame shifts (incorporation of nucleotides in each cycle is not 100% complete) and positive frame shifts (the population of nucleotides that is not fully degraded by the apyrase and can therefore be incorporated after the next nucleotide) that eventually will generate high levels of noise. Approximately 1.2 million wells will give one unique sequence of 400 bp, on average generating less than 500 million bases (Mb) in one single run. Whole-genome sequencing has been performed on bacterial genomes in single runs. An oversampling of 20× permits the identification of PCR-introduced errors and to call homopolymeric errors.

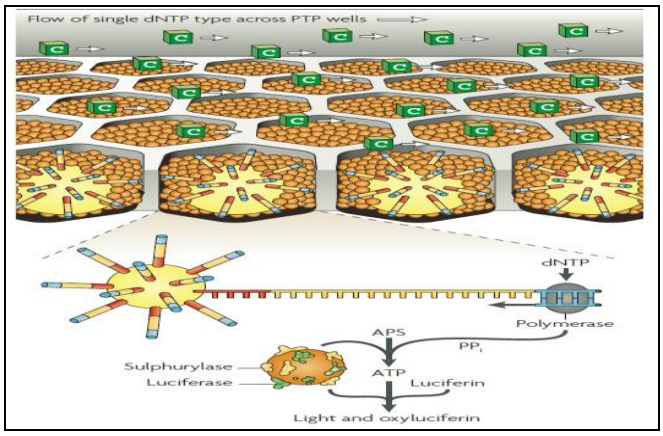

Fig 4: Basic principle behind Pyrosequencing

7. Illumina/Solexa sequencing

The Illumina/Solexa approach achieves cloning-free DNA amplification by attaching single-stranded DNA fragments to a solid surface known as a single-molecule array, or flowcell, and conducting solid-phase bridge amplification of single-molecule DNA templates. In this process, one end of single DNA molecule is attached to a solid surface using an adapter; the molecules subsequently bend over and hybridize to complementary adapters (creating the "bridge"), thereby forming the template for the synthesis of their complementary strands. After the amplification step, a flow cell with more than 40 million clusters is produced, wherein each cluster is composed of approximately 1000 clonal copies of a single template molecule. The templates are sequenced in a massively parallel fashion using a DNA sequencing-by-synthesis approach that employs reversible terminators with removable fluorescent moieties and special DNA polymerases that can incorporate these terminators into growing oligonucleotide chains [7]. The terminators are labeled with fluors of four different colors to distinguish among the different bases at the given sequence position and the template sequence of each cluster is deduced by reading off the color at each successive nucleotide addition step. Although the Illumina approach is more effective at sequencing homopolymeric stretches than pyrosequencing, it produces shorter sequence reads and hence cannot resolve short sequence repeats. In addition, due to the use of modified DNA polymerases and reversible terminators, substitution errors have been noted in Illumina sequencing data. Typically, the 1G-genome analyzer from Illumina, Inc. is capable of generating 35-bp reads and producing at least 1 GB of sequence per run in 2-3 days. The raw accuracy is said to be at 98.5% and the consensus (3×coverage) at 99.99%. The cost per base is approximately 1% of the cost for Sanger sequencing [8].

8. ABI Solid Sequencing

The ABI Solid sequencing system (2007) is based upon sequencing by ligation [9]. The ABI Solid sequencing starts with the basic step i.e. fragmentation of the target DNA

generation of library. The fragments are ligated to adapters and then bound to beads. The emulsion PCR amplifies DNA fragments on the beads. A water droplet in oil emulsion contains the amplification reagents and only one fragment bound per bead. During amplification at first, a primer is hybridized to the adapter and then specially designed probe (different fluorescently dyed oligonucleotide 8-mer probe) is attached to the primer by ligase enzyme along with the addition of ligation mixture. In these octamers, the first two are the actual bases, next three are universal bases and last three are universal bases with fluorescent dye, measures for fluorescence and cleaved in each cycle, so attached probe is only 5 nucleotide long i.e. fluorescence measurement for every fifth base. The ligated octamer oligonucleotides are cleaved off after the fifth base, removing the fluorescent label, then hybridization and ligation cycles are repeated, this time determining bases 9 and 10 in the sequence; in the subsequent cycle bases 14 and 15 are determined, and so on (Fig: 5). The error rate with this technique is very low along with a great output of 3 to 10 Gb per run.

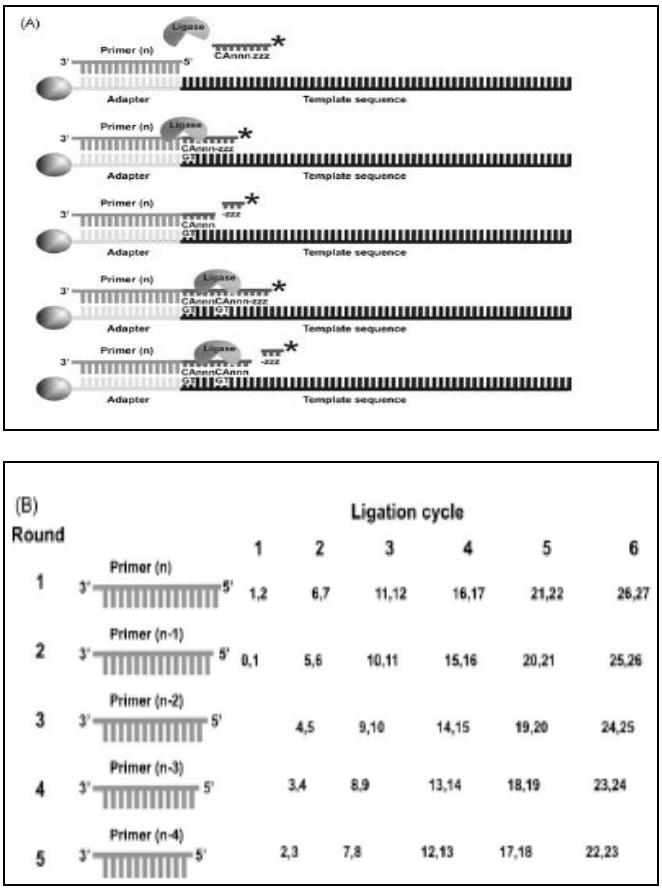

Fig 5: Steps involved in Solid sequencing

9. Ion torrent technology

For the PGM/ Proton sequence platforms the sequence templates are generated on a bead or a sphere via emulsion PCR. An oil–water emulsion is created to partition small reaction vesicles that each ideally contains one sphere, one library molecule and all the reagents needed for amplification. Two primers that are complementary to the sequence library adapters are present, but one is only present in solution while the other is bound to the sphere. This serves to select for the library molecules with both an A and a B adapter while excluding those molecules with two A or B adapters from loading on the beads during emPCR. In addition, this ensures a uniform orientation of the sequence library molecules on the sphere. During the emPCR steps, individual library molecules get amplified to millions of identical copies that are bound to the beads to allow ultimate detection of the signal.

Although one emPCR reaction can generate billions of template spheres, some aspects inherent to the emPCR method, in addition to the general biases during PCR amplification, prevent optimal output. Due to the double Poisson distribution behavior, it is impossible to achieve optimal loading of one library molecule into all individual vesicles. In fact 1/3 of the vesicles will have the one molecule to one vesicle ratio, the remaining 2/3 will be either without a molecule or have more than one. In addition, breaking of the emulsion and recovery of the spheres are inefficient even with the latest automated systems. In the final step spheres containing amplified DNA are selected in an enrichment step from empty spheres and the loaded spheres are deposited into the sequencing chip.

The Ion torrent chip consists of a flow compartment and solid-state pH sensor micro-arrayed wells that are manufactured using processes built on standard CMOS technology. The detection of the incorporated bases during sequencing is not based on imaging of fluorescent signals but on the release of an H+ during extension of each nucleotide. The release of H+ is detected as a change in the pH within the sensor wells. Due to the lack of the time consuming imaging a sequencing run can be completed within 4 h. Since there is no detectable difference for H+ released from an A, C, G or T bases, the individual dNTPs are applied in multiple cycles of consecutive order. If upon delivery of a dNTP no change in pH is detected in a specific well, that nucleotide is not present in the template at the next available position. Alternatively, if a change in pH is detected, that base is in the template. In contrast to the Illumina's SBS method the dNTPs used are not blocked and when the template contains a series of a nucleotide after each other (a homopolymer stretch), the entire stretch of identical bases will be extended, leading to an accordingly stronger pH change which is directly proportional to the number of identical bases incorporated. Relative to a single 'A' a stretch of 'AA' will give a 2 fold increase in the pH, while an 'AAA' template will yield a 1.5 fold (3/2) increase in pH relative to 'AA' and for 6 vs. 5 identical bases this relative increase is just 1.2 fold. This decrease of the relative increase of the change in pH as the homopolymer length increases reduces the probability by which a homopolymer region is called correctly.

The dNTPs are added in a predefined flow order. At the first release of the system, this order was a repetitive T–A–C–G sequence. As with Illumina sequencing, not all of the template molecules on a template sphere get extended in perfect synchrony. On average 0.5–1% of the molecules deviate from the flow either because they lag behind due to improper extension or they advance ahead due to carry over of dNTPs from a previous cycle (Fig: 6). In order to minimize this de-phasing, the flow order was changed to a more sophisticated sequence that incorporated A–T–A catch-up type flows. This scheme allows incomplete extension of the A nucleotide to catch up after the T base. Although this does come at a cost of decreased overall read length, the overall quality of the read does improve. Still, the quality of the reads gradually decreases towards the ends of the reads. By taking into account the flow order, it is possible to make flow-aware base caller algorithms and flow-space aware aligner software and variant detection tools that take the actual flow order into account when processing the data in order to generate higher accuracy data. The present error rate for substitutions is

~0.1% which is very similar to that of the Illumina systems. The main point of criticism the system endures is the homo-polymer errors. Despite many improvements the 5-mer homo-polymer error rate is still at 3.5%.

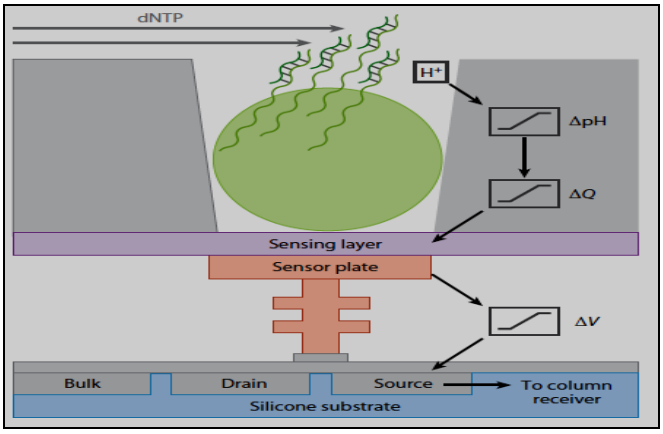

Fig 6: Mechanism involved in Ion torrent sequencing

Since the initial release of the Ion torrent platform, this technology has evolved at a very rapid pace. Through increasing the total surface area of the chips and the sensor well density the average read length from 100 up to 400 bp could be increased. The Proton-I chips currently yield 60-80 million reads per run, reaching 10 GB. This is enough to sequence two human exomes at ~50× coverage.

10. Helicos true single molecule sequencing

The concept of sequencing-by-synthesis without a prior amplification step i.e. single-molecule sequencing is currently pursued by a number of companies [10]. Helicos Biosciences has an instrument, the HeliScope™, with a claimed throughput of 1.1 Gpb per day. Single fragments are labeled with Cy3 for localization of template strands on an array and a predefined, Cy5-labeled nucleotide (for instance "A") are incorporated, detected by a fluorescent microscope and cleaved off in each cycle. Four cycles, one for each nucleotide, constitutes a "quad" and multiple quad runs are claimed to produce read lengths of up to ~55 bases. At 20 bases or longer, 86% of the strands are available and at 30 bases, around 50%. True single-molecule sequencing (tSMS™) is achieved by initially adding a poly A sequence to the 3′-end of each fragment, which allows hybridization to complementary poly T sequences in a flow cell. After hybridization, the poly T sequence is extended and a complementary sequence is generated. In addition, the template is fluorescently labeled at the 3′-end and thus, illumination of the surface reveals the location of each hybridized template. This process allows generation of a map of the single molecule landscape before the labeled template is removed. Fluorescently labeled nucleotides are added, one in each cycle, followed by imaging. A cleavage step removes the fluorophore and permits nucleotide incorporation in the next round.

A different, although very promising, approach is taken by Pacific Biosciences. The technology, denoted Single-molecule Real Time Sequencing-by-synthesis (SMRT shown read lengths of single DNA fragments of over 1500 bases in 3000 parallel reactions. The heart of the technology is so called zero-mode waveguides (ZMW), which essentially are nanometer scale wells with a diameter of 70 nm [11]. Light bulges inward at the opening, permitting illumination of a detection volume of 20 zl where a single DNA polymerase is immobilized. Nucleotides, fluorescently labeled at the terminal phosphate, are incorporated by the polymerase and thereby exposing its base-specific fluorophore for a few milliseconds which is enough for detection. Benefits are long read lengths of thousands of bases in one stretch and high speed (10 bases per second and molecule). It is still at the proof-of-concept stage and no commercial instrument is ready. Thousands of ZMWs in parallel may in a future instrument (no sooner than 2010) generate 100 gigabases per hour. A second-generation instrument capable of sequencing a human genome for $1000 is an additional number of years in the future.

11. Nano pore sequencing

Nanopore-based sequencers, as the fourth-generation DNA sequencing technology, have the potential to quickly and reliably sequence the entire human genome for less than $1000, and possibly for even less than $100. The nanopore approach will be one option for the fourth-generation low-cost and rapid DNA sequencing technology. Nanopore-based technologies originated from the Coulter counter and ion channels. With the application of an external voltage, particles with sizes slightly smaller than the pore size are passed through the nanometer-sized pores are either embedded in a biological membrane or formed in solid-state film, which separates the reservoirs containing conductive electrolytes into cis and trans compartments. When the pore is blocked by an analyte, such as a negatively charged DNA molecule added into the cis chamber, current flowing through the nano pore would be blocked, interrupting the current signal (Fig: 7).

The physical and chemical properties of the target molecules can be calculated by statistically analyzing the amplitude and duration of transient current blockades from translocation events. Nanopores as single-molecule sensing technologies have great potential applications in many areas, such as analysis of ions, DNA, RNA, peptides, proteins, drugs, polymers, and macromolecules, as previously reviewed.

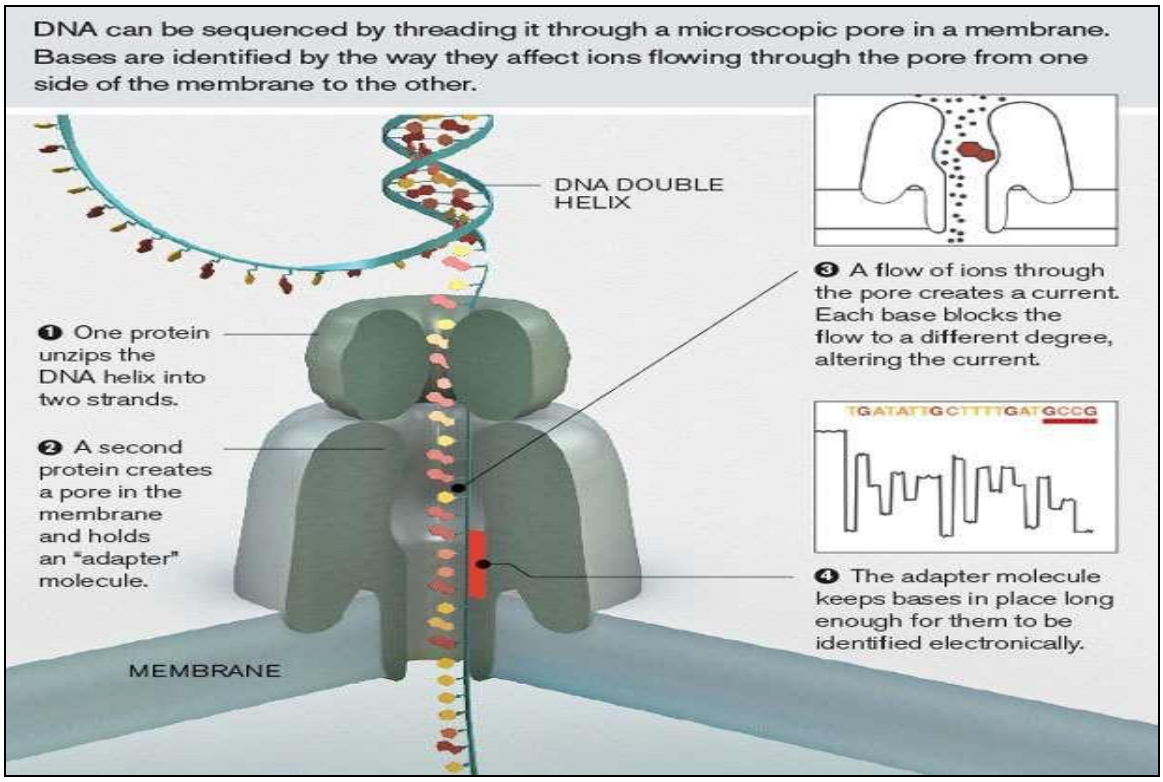

Fig 7: Mechanism involved in Nanopore sequencing

12. Application of Next generation sequencing in animal genetics

12.1 Transcriptome sequencing

The sequencing of cDNA rather than genomic DNA focuses analysis on the transcribed portion of the genome. This focus reduces the size of the sequencing target space, which can be viewed as desirable given the fact that, even with next-generation sequencers, sequencing an entire vertebrate genome is still an expensive undertaking. Transcriptome sequencing has been used for Gene expression profiling, Genome annotation, Rearrangement detection that causes aberrant transcription events, noncoding RNA discovery and quantification [12].

12.2 Analysis of epigenetic modifications of histones and DNA

Epigenetics is the study of heritable gene regulation that does not involve the DNA sequence. The two major types of epigenetic modifications regulating gene expression are DNA methylation by covalent modification of cytosine-5′ and posttranslational modifications of histone tails there are three main approaches to detecting DNA methylation on a large scale, including restriction endonuclease digestion coupled to microarray technology, bisulfite sequencing, and immunoprecipitation of 5′-methylcytosine to separate methylated from unmethylated DNA. Bisulfite sequencing is based on the chemical property of bisulfite to induce the conversion of cytosine residues to uracils while leaving 5′ Œ-methylcytosines intact. Therefore, sequencing of bisulfite treated DNA will reveal the positions of methylated cytosines (those positions that remained cytosines following the treatment) [13].

12.3 Genomic selection

Advent of all these sequencing techniques of sequencing forms the basis for the Genomic Selection. Decreased cost of sequencing triggered the sequencing of all the animal's genomes, which results in the identification of tens of thousands of SNPs present in the genome.

12.4 Identification of all mutation in an organism at genomic level

12.5 Nucleotide variation profiling

12.6 Selection of resistant animal to disease

12.7 Metagenomics

Using brute force sequencing, simply reading all DNA sequences present in a sample, metagenomics is a way to make an inventory of what is present in a sample, of what is living where. A simple but effective application of this is trying to detect the cause of an infectious disease. Simply analyzing all DNA from control versus infected (diseased) individuals will reveal the "extra" DNA, which most likely derives from the infectious agent. The approach was used successfully to identify e.g. colony collapse disorder killing honeybees but also to identify the cause of diseases that killed thousands of humans in the past, inclusive the Black Death.

Metagenomics can be performed by a sequence it all approach or by focusing on specific uniformly conserved sequences like e.g. ribosomal RNA genes only [14]. The latter approach has two main advantages; the complexity of the data obtained is much smaller and more sequences can be assigned to a specific organism or a group of related organisms. The latter facilitates some semi-quantitative analysis which is much more difficult when analyzing all sequences mixed from many organisms with largely varying genome sizes. Metagenomics focussing on rDNA genes has been used to study many different things, incl. e.g. the effect of the 2004 tsunami on microbial ecologies in marine, brackish, freshwater and terrestrial communities in Thailand [15].

12.8 Whole-exome sequencing (WES)

Only the coding regions of the genome are sequenced. The exome represents less than 2% of the human genome, but contains ~85% of known disease-causing variants. For this reason, WES has been extensively used for clinical studies in the recent years, and is giving rise to promising novel

diagnostic tools that have the potential to transform medical healthcare in the near future. For example, an experimental approach for comprehensive WES of circulating tumor cells from cancer patients has recently been described [16].

12.9 Single cell genomics

An exciting new field is a single cell genomics. A major objective of this field is to reconstruct cell lineage trees using somatic mutations that arise due to DNA replication errors. As a result, each cell in a multicellular organism carries a genomic signature that is probably unique. Cell lineage trees provide important information and have applications in developmental biology and tumor biology [17]. For example, sequencing the genomic DNA of individual breast cancer cells allowed researchers to reconstruct the tumor population structure and evolutionary history [18].

12.10 Large-scale identification and development of molecular markers

One of the most important application of NGS within ecological and population genetics is the development of molecular markers on a large scale. NGS generally allows for cost-effective and rapid identification of hundreds of microsatellite loci and thousands of SNPs, even if only a fraction of a sequencing run is used. This will, for example, facilitate QTL mapping studies and will increase the quality of outlier- and structure analyses. Massively increasing the number of markers will enable researchers not only to get better precision in population genetic studies (QTL and linkage disequilibrium (LD) mapping projects and kinship assignments but also to pursue topics such as historical demographic patterns, introgression and admixture.

13. Conclusions

The short read structure of next-generation sequencers provides potential problems for sequence assembly particularly in areas associated with sequence repeats. However, it has found broad applicability in sequence census studies, wherein determining the sequence of the whole DNA molecule is not essential. The short read length also necessitates the development of paired-end sequencing approaches for improved mapping efficiency. To date, such approaches have been reported for the 454 technology and are being made available for the Illumina and Solid platforms. The accuracy of next-generation sequencers is improving, but users generally rely on relatively high redundancy of sequence coverage to determine reliably the sequence of a region, particularly of that containing a polymorphism. Addressing the accuracy issue by improving the reaction chemistry has the potential of further decreasing the current sequencing cost associated with next-generation sequencers.

Next-generation sequencing technologies have found broad applicability in functional genomics research. Their applications in the field have included gene expression profiling, genome annotation, small ncRNA discovery and profiling, and detection of aberrant transcription, which are areas that have been previously dominated by microarrays. Significantly, several studies found that 454 sequencing correlated well with the established gene expression profiling technologies such as microarray results (correlation coefficients of (0.83-0.91) and moderately with SAGE data (correlation coefficients of 0.45). While the transcriptome sequencing studies discussed here predominantly used the 454 technology, Illumina and Solid technologies also offer significant potential for such applications. Another major functional genomics application is determining DNA sequences associated with epigenetic modifications of histones and DNA. Next-generation sequencing approaches have been used in this field to profile DNA methylations, posttranslational modifications of histones, and nucleosome positions on a genome-wide scale. While Sanger sequencing has previously addressed these areas, next-generation technologies have improved upon the throughput, the depth of coverage, and the resolution of Sanger sequencing studies.

Despite the recent exciting research advances involving next generation sequencers, it should be noted that method development is still in its infancy [19]. Efficient data analysis pipelines are required for many applications before they become routine, and more studies are needed to address the robustness of these techniques as well as the correspondence of results with those obtained by previous methods. Although next-generation sequencers are already being widely used, there are other sequencing methods, such as nanopore sequencing, whose scalability is being explored to decrease the sequencing cost and enhance throughput even further [20].

14. References

1. Sanger F, Donelson JE, Coulson AR, Kossel H, Fischer D. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. Proc. Natl. Acad. Sci. USA. 1973; 70:1209-13.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 1977; 74:5463-67.
3. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol. Biol. 1975; 94:441-48.
4. Sanger F, Brownlee GG, Barrell BG. A two-dimensional fractionation procedure for radioactive nucleotides. J Mol. Biol. 1965; 13:373-98.
5. Metzker ML. Sequencing technologies-the next generation. Nat. Rev. Genet. 2009; 11:31-46.
6. Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005; 15:1767-1776.
7. Rogers YH and Venter JC. Genomics: massively parallel sequencing. Nature. 2005; 437:326-327.
8. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R *et al*. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc.Natl Acad. Sci. USA. 2006; 103:11240-11245.
9. Gait MJ, Matthes HW, Singh M, Sproat BS, Titmas RC. Rapid synthesis of oligodeoxyribonucleotides. Solid phase synthesis of oligodeoxyribonucleotides by a continuous flow phosphotriester method on a kieselguhr-polyamide support. Nucleic Acids Res. 1982; 10:6243-54.
10. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc. Natl Acad. Sci. USA. 2003; 100:3960-3964.
11. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG. Webb WW. Zeromode waveguides for single-molecule analysis at high concentrations, Science. 2003; 299:682.686.
12. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1):57-63.
13. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA *et al*. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat

Methods. 2010; 7(6):461-465.
14. Abubucker S, Segata N, Goll J, Schubert AM, Izard J. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput. Biol. 2012; 8:e1002358
15. Somboonna N, Wilantho A, Jankaew K, Assawamakin A, Sangsrakru D. Microbial ecology of Thailand tsunami and nontsunami affected terrestrials. Plos One. 2014; 9:e94236.
16. Lohr JG. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nat. Biotechnol. 2014; 32:479-484.
17. Reizel Y. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. Plos Genet. 2011; 7:e1002192.
18. Navin N. Tumour evolution inferred by single-cell sequencing. Nature Pettersson, Erik; Lundeberg, Joakim and Ahmadian, Afshin. Generations of sequencing technologies Genomics. 2009; 93(472): 90-94, 105-111.
19. Chan EY. Advances in sequencing technology. Mutat. Res. 2005; 573:13-40.
20. Buermans HPJ, Dunnen JT. Next generation sequencing technology: Advances and applications. Biochimica ET Biophysica Acta. 2014; 1842:1932-1941.