REVIEW

Next-Generation Sequencing and its new possibilities in medicine

Marianna Nagymihály[1], Attila Szűcs[1], Attila Kereszt[1,2]*

[1]Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary
[2]Seqomics Biotechnology Ltd, Mórahalom, Hungary

ABSTRACT Next-Generation Sequencing (NGS) originally refers to high-throughput, massively parallel sequencing methods that allow the sequencing of up to billions of small (50-1000 bp), amplified DNA fragments at the same time but nowadays, there are NGS techniques that determine the sequence of long (up to 50 kbp) single molecules. Over the past years, NGS technologies become widely available with increasing throughput and decreasing sequencing costs per base making them more cost effective than the previously used capillary sequencing methods based on Sanger biochemistry. Nowadays, high-throughput DNA sequencing is routinely used on a wide range of important fields of biology and medicine enabling large-scale sequencing projects like analysis of complete genomes, disease association studies, whole transcriptomes, methylomes and provide new insights into complex biological systems. In addition, more and more NGS-based diagnostic tools are being introduced into the clinical practice, for example, on the fields of oncology, inherited and infectious diseases or pre-implantation and prenatal genetic screenings. **Acta Biol Szeged 59(Suppl.2):323-339 (2015)**

KEY WORDS
genome
high-throughput
massively parallel sequencing
single molecule

The first-generation DNA sequencing

Although the structure of DNA was described in 1953 (Watson and Crick 1953), it took more than 20 years to develop the first efficient method for sequencing DNA (Sanger and Coulson 1975). Before this work, several ad hoc methods involving RNA synthesis and enzymatic digestion were used to determine some nucleotide sequences. In 1971, Wu was able to determine 12 nucleotides in the single-stranded cohesive ends of bacteriophage λ DNA. This method used radioactively labelled nucleotides, DNA polymerase, complex nucleotide digestion and chromatography (Wu and Taylor 1971). In 1973, Gilbert and Maxam published a 24-bp sequence from *Escherichia coli*, by a method called wandering-spot analysis (Gilbert and Maxam 1973). Sanger used a short synthetic oligonucleotide primer, which hybridized to a single-stranded DNA template, DNA polymerase, and four radioactively labelled deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, and dTTP) to generate newly synthetized DNA fragments of different sizes. The sample was subdivided into 4 different reaction mixtures; all of them contained only three of the four nucleotides, so the sequencing reactions were terminated at the missing nucleotides. After denaturation, the synthetized DNA fragments in the four samples were size-separated in polyacrylamide gel, and the sequence could be determined from radioautograph. This method was the first technique which allowed the sequencing of long DNA sequences and could be applied to any DNA molecule. The disadvantages of this method were the requirement for prior knowledge on the DNA sequence to which oligonucleotide primers would hybridize and the uneven distribution of the produced fragments with the desired lengths.

In 1977, dideoxynucleotide-triphosphates (ddNTPs lacking the 3' hydroxyl group) were introduced as "chain terminators" (Sanger et al. 1977) and this first-generation technology dominated the DNA sequencing field for more than 30 years. In this approach, four separate sequencing reactions, each containing the four dNTPs (one of them was radioactively labelled) and one of the ddNTPs were set up and whenever a dideoxynucleotide was incorporated into the DNA by the DNA polymerase, the chain elongation stopped, thus, a mixture of truncated fragments of varying lengths, all starting at the same primer and ending with the same base was produced. The fragments were size separated by poly-acrylamide gel electrophoresis at one nucleotide resolution and the sequence was deduced from the autoradiograms. The last major improvement was when fluorescent labelling replaced

Submitted April 20, 2015; Accepted Sept 10, 2015
*Corresponding author. E-mail: kereszt.attila@brc.mta.hu

the use of the expensive and hazardous radioactively labelled nucleotides (Smith et al. 1986). Based on the detection of the four fluorescent dyes that were used for the labelling of the four ddNTPs was the first commercial semiautomatic DNA sequencer developed by Applied Biosystems Inc. (ABI). This technology allowed sequencing in a single reaction. An incremental improvement was when capillary tubes replaced polyacrylamide gels which allowed more consistent results and also increased the speed of electrophoresis (Drossman et al. 1990). This first-generation sequencing technology provided nearly all of the data for the Human Genome Project (International Human Genome Sequencing 2001; Venter et al. 2001) which took more than 10 years and cost more than 2.5 billion US$.

Second-generation sequencing technologies

Commercially available second-generation sequencing technologies include the MiSeq, NextSeq, HiSeq systems of Illumina that use reversible dye terminator technique, the sequencing by ligation (SOLiD) and the semiconductor (Ion Torrent) technologies of Life Technologies as well as the pyrosequencing based instruments (GS FLX, Junior) of Roche 454 (Metzker 2010). Although the sequencing biochemistry used by these different sequencing platforms is diverse, the library preparation workflow is quite similar.

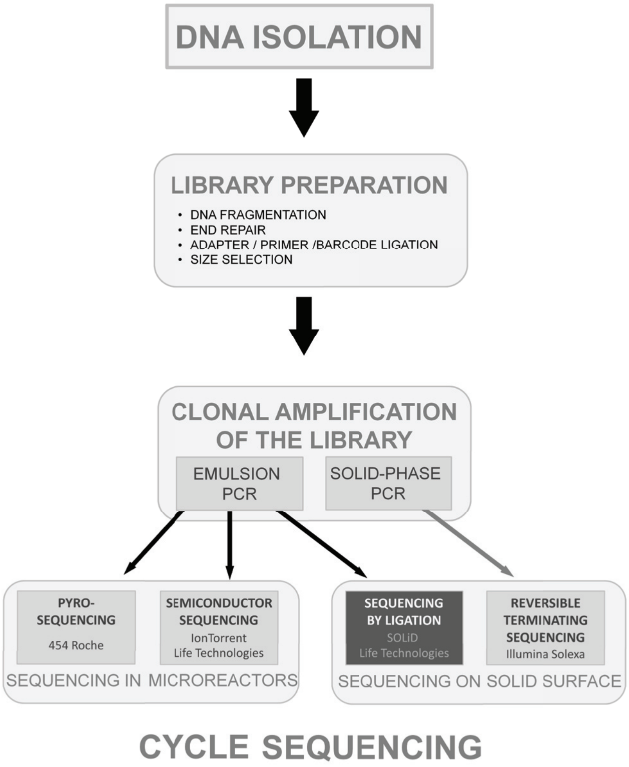

Figure 1. An overview of the steps of second-generation sequencing. The basic steps of the process are the same for each technology, diversification can be observed at the clonal amplification and cycle sequencing steps.

There are fundamental differences between the first- and second-generation technologies:

(i) Maximum 96 clones can be sequenced at the same time in the first-generation instruments while the sequence from up to billions of DNA fragments are determined in the second-generation sequencers;

(ii) One of the limitations of the first-generation sequencing technologies in performing genome-wide investigations (such as genome or transcriptome sequencing) is the requirement for the cloning and thus, the amplification of the target DNA fragments in bacterial hosts. This cloning step makes large scale sequencing projects manpower intensive and expensive, introduces biases and restricts the parallelism (Hall 2007). In contrast, all second-generation technologies directly use mechanically or enzymatically fragmented nucleic acids to prepare sequencing libraries by adding forward and reverse sequencing adaptors (Fig. 1). The DNA fragments (genomic or complementary), *i.e.* the sequencing libraries are then clonally amplified with the use of a single adaptor specific universal primer pair on a surface of streptavidin-coated DNA capture beads (Roche 454, SOLiD, Ion Torrent) or by 'bridge' PCR on a solid array surface used by the Solexa technology (Illumina). The clonal amplification method, first used by the Roche 454 technology (Margulies et al. 2005), circumvented the cloning requirement but it also can insert bias into the results.

(iii) Whereas individual samples (sequencing reactions) in first-generation sequencers is placed into lanes/capillaries and the sequence is determined by separating bases in space, second-generation sequencers rely on a huge number of nano-reactors containing amplified clonal clusters of DNA fragments where the sequence of bases is resolved in time by implementing cyclic-array sequencing (Jarvie 2005). The process is cyclic because in each sequencing cycle a single nucleotide or dinucleotide is investigated by an enzymatic process (synthesis or hybridization-ligation) to identify nucleotides in the template in a highly parallel way (Fig. 2 and 3).

The second-generation sequencing library preparation workflows are quite similar for the different sequencing platforms while the sequencing biochemistries described in detail at the reviewing of the different technologies are diverse. These systems are capable to sequence samples from a wide variety of starting materials including genomic DNA, PCR products, yeast and bacterial artificial chromosomes (YACs, BACs), as well as cDNA or RNA. In general, DNA samples must be fragmented usually into small pieces (100-800 bp) or

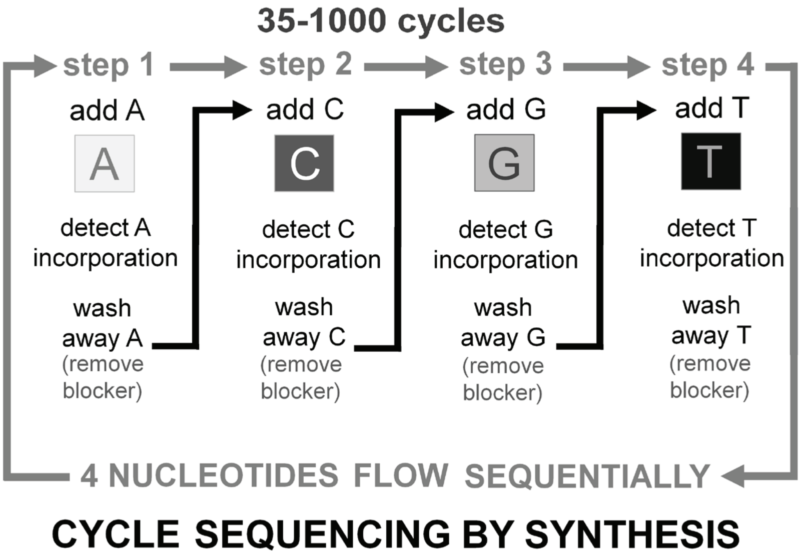
35-1000 cycles
step 1
step 2
step 3
step 4
add A
A
add C
C
add G
G
add T
T
detect A incorporation
detect C incorporation
detect G incorporation
detect T incorporation
wash away A (remove blocker)
wash away C (remove blocker)
wash away G (remove blocker)
wash away T (remove blocker)
4 NUCLEOTIDES FLOW SEQUENTIALLY
CYCLE SEQUENCING BY SYNTHESIS

Figure 2. General steps of cycle sequencing by synthesis. In one step of a cycle, microreactors and chips containing the clonally amplified library fragments are flown with a (labeled) nucleotide followed by the detection of nucleotide incorporation. Next, the non-incorporated nucleotides and blockers are washed away. There four steps of a cycle corresponds to the four nucleotides of DNA.

larger (2-20 kbp) fragments for a special application, the mate pair library preparation for *de novo* genome sequencing (Fig. 2). Adaptors (small nucleotide DNA sequences) are added to the ends of sample DNA that are necessary for subsequent steps including library amplification and quantification as well as sequencing. Specifically designed, oligonucleotide covered DNA Capture Beads (Roche 454, SOLiD, Ion Torrent) or glass surface (Illumina) are used to immobilize single stranded library fragments via hybridization to one (Roche 454, SOLiD, Ion Torrent) or both adaptors (Illumina). The ratio of the beads/surface and the fragments is adjusted in such a way that only one library fragment is bound on the majority of the beads (Roche 454, SOLiD, Ion Torrent) or the distance between the bound DNA fragments on the glass surface is larger than the distance between the oligonucleotides holding a single fragment (Illumina), thus these ‘bridges’ do not cross each other. The bead-bound library is emulsified with amplification reagents in a water-in-oil mixture resulting in micro-reactors containing just one bead with one unique library fragment. Clonal amplification occurs in these mini-reactors and the process is called emulsion PCR (emPCR), without contamination with other micro-reactors and DNA sequences. Similarly, the bound DNA fragments on the glass surface are amplified to a few hundred copies to form a cluster of identical DNA pieces. Beads from the emPCR are then enriched for template-positive beads by hybridization with capture beads and transferred either into micro-wells (Roche 454, Ion Torrent) or chemically cross-linked to a glass slide (SOLiD). At the end, up to billions of parallel sequencing reactions are carried out at the same time using sequencing primers hybridized to universal adaptors at specific position and orientation maintaining strand specific information.

Roche 454 pyrosequencing technology

The pyrosequencing is based on a sequencing-by-synthesis method which allows sequencing of a single DNA strand by synthesizing the complementary DNA strand base by base from an oligonucleotide primer and detecting which base was actually incorporated into the DNA strand at each step (Metzker 2010). This technique uses four enzymes (DNA polymerase, ATP sulfurylase, luciferase and apyrase) and their substrates (dNTPs, adenosine-5´-phosphosulfate (APS), and luciferin). If the tested nucleoside triphosphate is incorporated by the sequencing polymerase, it results in the release of inorganic pyrophosphate (PPi) which is used for ATP synthesis by the ATP sulfurylase in the presence of APS. The production of ATP leads to the generation of a light burst that accompanies the ATP-dependent conversion of luciferin to oxiluciferin by the activity of the luciferase enzyme. This chemiluminescence of each picotiter well is detected and measured by a CCD camera and analyzed by the software. The apyrase degrades the unincorporated nucleotides and ATP and the next sequencing cycle can be started by adding the next nucleoside triphosphate. The sequence of the DNA template is determined from a “pyrogram” which shows the light intensities detected during the sequencing cycles. Theoretically, the intensity of the chemiluminescent light

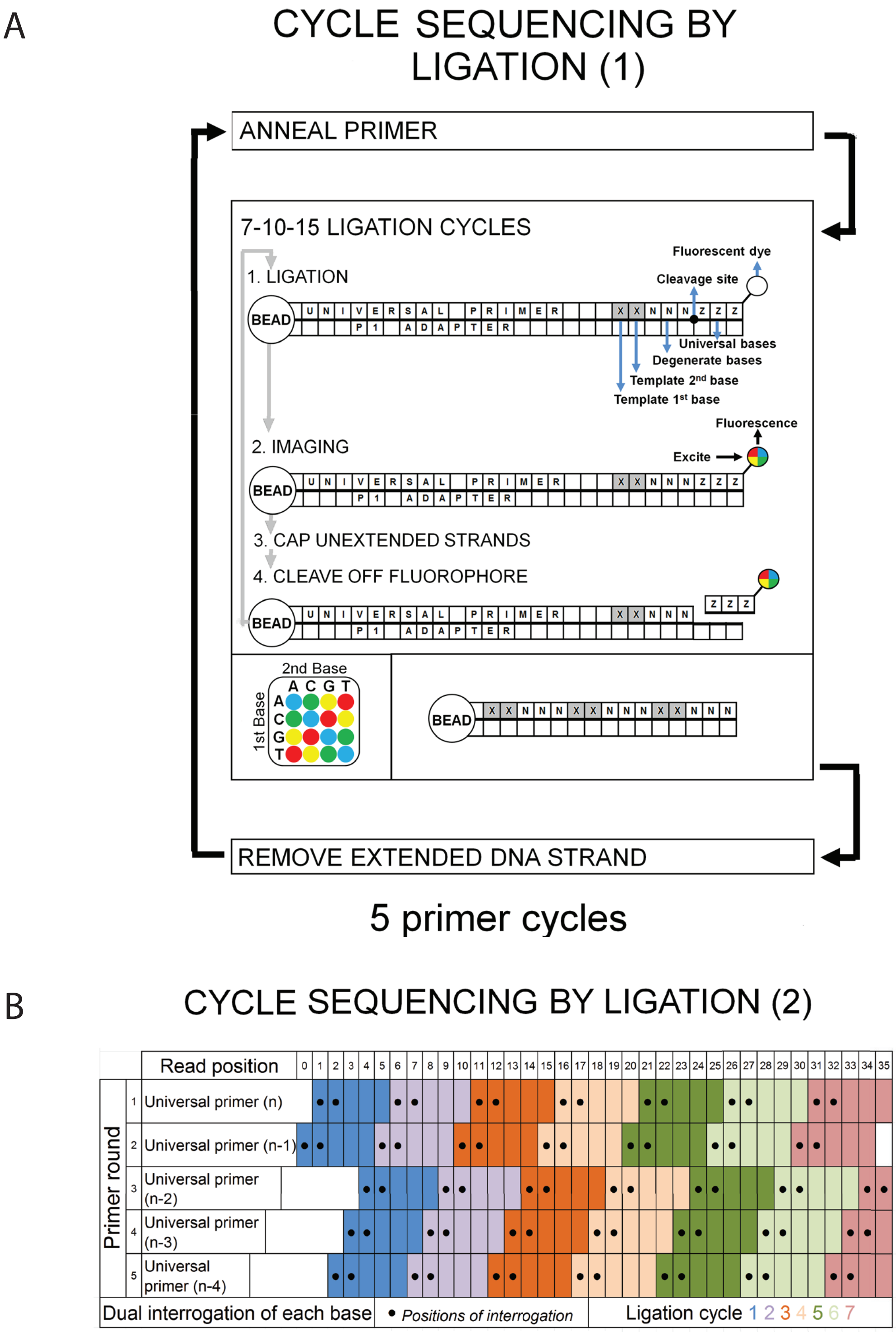

Figure 3. General steps (A) of cycle sequencing by ligation and (B) the interrogated bases in the different ligation cycles of the five primer cycles. (A) In the first step of a ligation cycle, a fluorescently labeled oligonucleotide probe is ligated to the sequencing primer (1st cycle) or to the extended second strand (2nd to 15th cycles). The colour of fluorescence depends on the first two bases (denoted by X) of the probe. Each colour label four dinucleotides. After capping of the unextended strands, last three nucleotides (denoted by Z) with the fluorophore are removed while three spacer nucleotides (denoted by N) retained in the growing DNA. After completion of the ligation cycles, the extended DNA strand is removed and another sequencing primer shifted with one nucleotide position is hybridized to the library fragment and ligation cycles begin. (B) Using this two-nucleotide encoding, each nucleotide is interrogated in two different primer cycles in combination with its preceding and following nucleotides.

produced should be proportional to the amount of the ATP, *i.e.* to the number of nucleotides incorporated from the tested one. However, the major limitation of the pyrosequencing approach is the homopolymeric sequence stretches because intensity of the generated light is easy to miscalibrate and the correct number of identical nucleotides cannot be determined. With the pyrosequencing technology a read length of 700 to 1000 bp can be achieved which was the key advantage of the Roche platforms. For certain applications like de novo whole genome sequencing or de novo transcriptomics these long reads have been critical. The desktop (Junior) and production-scale (FLX) systems of Roche are capable of generating 70 and 700 Mb of sequence data, respectively. Roche recently announced the discontinuation of these platforms

Life Technologies Ion Torrent instrument: semiconductor sequencing

This "sequencing by synthesis" technology differs from the other second-generation sequencing technologies in that no modified nucleotides, no optics and no cameras are used. Ion semiconductor sequencing is based on the detection of hydrogen ions that are released as by-products during nucleotide incorporation into the DNA strand by the polymerase (Metzker 2010). During each sequencing cycle the wells (which are basically miniature pH meters) of the semiconductor chip containing DNA template-positive beads are sequentially flooded with a single species of the four different (A, T, G, C) deoxyribonucleotide triphosphates. If the introduced deoxyribonucleotide triphosphates incorporated into the growing template strand, the released proton changes the pH of the solution which can be detected by the proprietary ion sensor of the corresponding well of the semiconductor chip. If two identical bases are incorporated into the DNA strand, the voltage will be double, and the chip will record two identical bases. If there is no deoxyribonucleotide triphosphate incorporation, no voltage change will be detected. In case of homopolymer repeats multiple deoxyribonucleotide triphosphate molecules will be incorporated in a single cycle and this leads to a corresponding number of released protons and a proportionally higher electronic signal. The unattached dNTP molecules are washed away before the next cycle and a different dNTP species is introduced. The major benefits of Ion Torrent semiconductor sequencing are rapid sequencing speed and low operating costs due to the avoidance of modified nucleotides and optical measurements. The system has the same limitation as pyrosequencing, *i.e.* the accurate detection of long homopolymer repeats. Currently, Life Technologies distributes two desktop sequencers of different throughput, the Personal Genome Machine (PGM) and the Proton which can produce up to 2 and 10 Gbp sequence data, respectively. The main advantages of the Ion Torrent systems are the short sequencing time (2-4 hours) and the scalability which is achieved with the use of chips that have different numbers of minireactors.

The reversible dye terminator sequencing: the Illumina/Solexa Technology

The Illumina/Solexa technology is unique among the second-generation sequencing methods that both clonal amplification of the library fragments and sequencing (also by synthesis) is performed on the solid surface of the flow cells (Metzker 2010). Because of the sequencing chemistry, *i.e.* the use of nucleotides labeled with four different cleavable fluorescent dyes and reversibly blocked on the 3' C-atom, the DNA polymerase can incorporate only a single nucleotide, thus, this system does not suffer from the presence of homopolymeric stretches. At present, the majority of the Illumina instruments support the sequencing of 2 x 125 or 2 x 150 nucleotides; however, the MiSeq machine is capable of reading 2 x 300 nucleotides in paired-end mode. The read lengths are limited because of signal decay and de-phasing caused by incomplete cleavage of fluorescent modification or the 3' terminating moiety. Currently, the portfolio of Illumina has the sequencing power for every scale: The desktop sequencers MiSeq and NextSeq can provide up to 15 and 120 Gbp sequence data per run, respectively, allowing targeted or even whole genome sequencing. Using the HiSeq series instruments, 4-6 billion reads, 1-1.8 Tbp sequence can be obtained in a single run that enable researchers to perform large-scale sequencing projects. The HiSeq X Ten system contains ten HiSeq X instruments and is designed for population-scale projects to sequence thousands of human genomes.

Sequencing by ligation: the SOLiD system of Life Technologies

Unlike the other three second-generation methods that are based on sequencing by synthesis, the SOLiD (Sequencing by Oligonucleotide Ligation and Detection) technology developed by Life Technologies uses hybridization-ligation cycles (Fig. 3) (Metzker 2010). In the first primer cycle, Universal seq primer (n) attaches to the P1 adaptor of the clonally amplified library fragments. During the first ligation cycle, 16 dinucleotide probes labelled by four different fluorescent dyes (each dye label four dinucleotides) and containing a three-nucleotide long non-specific spacer are used to interrogate the 1st and 2nd positions of the fragment and the probe with complementary dinucleotide is ligated to the primer. After imaging of the fluorescence, the unextended strands are capped with the same mixture of nonfluorescent probes and treated with phosphatase to prevent any unextended strands for contributing to "out of phase" ligation events. Finally, the

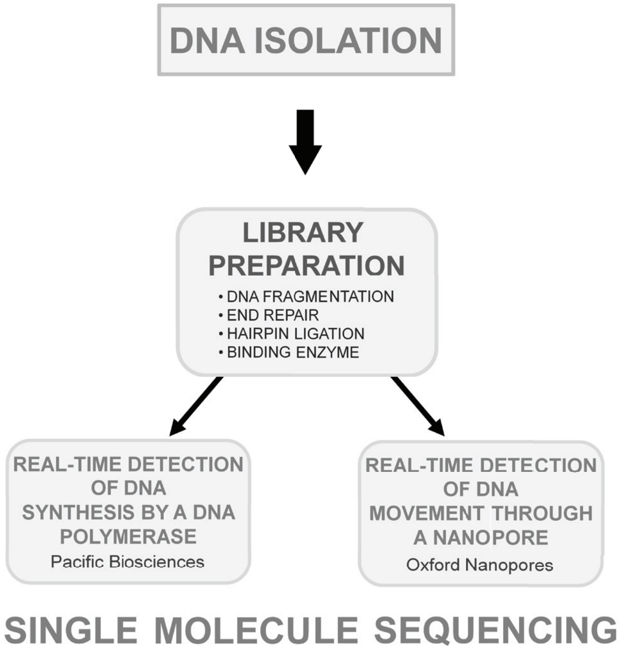
DNA ISOLATION
LIBRARY PREPARATION
• DNA FRAGMENTATION
• END REPAIR
• HAIRPIN LIGATION
• BINDING ENZYME
REAL-TIME DETECTION OF DNA SYNTHESIS BY A DNA POLYMERASE
Pacific Biosciences
REAL-TIME DETECTION OF DNA MOVEMENT THROUGH A NANOPORE
Oxford Nanopores
SINGLE MOLECULE SEQUENCING

Figure 4. An overview of the steps of third-generation sequencing. Single molecule sequencing does not require PCR amplification after library preparation.

fluorescent label is removed and the second ligation cycle to interrogate the 6th and 7th position of the fragment may start. After 7, 10 or 15 ligation cycles are performed, the extension product is removed and the template is reset with a primer complementary to the n-1 position for the second round of ligation cycles. Altogether, five primer cycles with the same number of ligation cycles are completed to obtain 35, 50 or 75 nucleotide long sequence reads from the fragments. Through the primer reset process, each base is interrogated in two independent ligation reactions by two different primers which enable the distinction between a sequencing error and a sequence polymorphism. The SOLiD 5500 Wildfire instrument is capable of producing 320 Gb of sequence data per run from 3.2 billion 2 x 50 bp paired-end reads.

Third-generation sequencing technologies

Despite the obvious advantages of the second-generation sequencing (such as extremely high throughput, low cost per nucleotide sequenced), there are some weaknesses related to these techniques. For example, the PCR steps during library preparation may introduce biases such as introducing "mutations" because of the lack of proof-reading or overrepresentation of certain fragments. The lower than 100% efficiency of the different steps during the alternating phases of nucleotide incorporation, signal detection and signal removal may lead to dephasing (reading not only the nth but also the (n-1)th nucleotide in the clonal population) which results in the decrease of read quality towards the end. One other major disadvantage of the second-generation sequencing techniques is the read length which is usually shorter than the one obtained in first-generation Sanger sequencing which is capable yielding routinely ~1000-1200 nucleotide long sequences. The longest second-generation reads are obtained by Roche 454 pyrosequencing are comparable to those of the first-generation sequencers (700-1000 nucleotide), however, the other systems produce much shorter reads of 35-75 (SOLiD), 300 (Illumina) and 400 (Ion Torrent) nucleotide length. Short reads cause difficulties for the assembly of genomes that contain repetitive sequences which is the case for most organisms. Recent improvements like mate-paired or paired-end sequencing make the "*de novo*" assembly of high-quality genomes possible. However, the genome of only a limited number of eukaryotic species has been completely sequenced due to the limitations of second-generation sequencing technologies including biases resulting from base compositions and the inability to sequence through large repeated elements.

On the rapidly growing field of next-generation sequencing technologies, the third generation platforms can overcome the limitations of the current sequencing techniques. Today, Pacific Biosciences and Oxford Nanopore provide single-molecule, real-time (SMRT) sequencing systems which achieve extra-long reads (up to 50 kbp) with high consensus accuracy that simplify and improve genome assembly. As the simplified library construction does not require amplification steps, no PCR biases and errors are introduced. In addition, as only one DNA molecule is read during the sequencing, errors from dephasing are not a problem. Though the accuracy of the raw reads are lower compared to data from second-generation sequencers, sequencing the same template molecule more than once and construction of consensus read sequences can provide high quality sequences. These technologies even allow the recognition of different base modifications (such as methylation on adenine and cytosine residues) directly without the use of chemical modifications, thus, allowing high-resolution mapping of the epigenome. In addition, RNA molecules without conversion into complementary DNA can be directly sequenced.

The library preparation for the third-generation sequencers includes the fragmentation and end-repairing of DNA which is followed by the ligation of hairpin sequences to both (Pacific Biosciences) or one (Oxford Nanopore) end of the fragments. Before starting the sequencing, an oligonucleotide primer is added (Pacific Biosciences) and a polymerase is bound to the DNA. Note that there is no PCR amplification step in this process that could introduce bias (Fig. 4).

Single molecule real-time sequencing

Single molecule real-time (SMRT) sequencing is a parallelized single molecule DNA "sequencing"-by-synthesis" technology developed by Pacific Biosciences (Metzker 2010). The technology based on direct observation of DNA synthesis of single DNA molecules in real-time in ~75000 zero-mode waveguides (ZMWs) placed onto a SMRT chip (Levene et al. 2003). The ZMW is cylindrical hole which is ~70 nm in diameter "drilled" into a 100 nm thin metal film supported by a transparent glass/silica substrate. As the wavelength of the laser light used for illumination is larger than the diameter of the ZMW, the light cannot pass through the ZMW's aperture, rather it decays exponentially inside the chamber (Metzker 2010) and only penetrates the lower 20-30 nm of the waveguide creating a detection volume of ~20 zeptoliters (2 x 10^{-21} liters). A single DNA polymerase molecule is attached to the bottom of each ZMW which synthesizes the DNA using 4 dNTPs that are labelled fluorescently with 4 different dyes on the γ-phosphate and not on the bases as the nucleotides used in first- and second-generation sequencing. The background signal from the fluorescently labelled nucleotides is very low because they diffuse through the illuminated part of the ZMW very quickly. However, when the DNA polymerase on the bottom of the ZMW incorporates a nucleotide, the enzyme holds it in the detection volume for tens of milliseconds which is orders of magnitude longer than the average time the diffusing nucleotides spend there. The presence of the nucleotide in the detection volume for so long creates a bright fluorescence pulse which can be distinguished from the background. After the nucleotide incorporated, the pyrophosphate with the fluorescent dye is released and quickly diffuses from the detection volume and the background level fluorescence is restored. The base call is made according to the corresponding fluorescence of the dye. Typical synthesis rates in SMRT sequencing are 1-3 bases per second (Korlach et al. 2010). The fluorescent signal is not only characterized by the emission spectra but also by its duration and the interval between the signals defined as pulse width and interpulse duration (IPD), respectively, which give information about the kinetics of the polymerase. The presence of modified nucleotides alters polymerase kinetics and therefore enables the detection of many types of DNA base modifications (*e.g.*, methylation or hydroxymethylation) which is one of the great advantages of this technology (Flusberg et al. 2010). Although the accuracy of the base calling using this approach is lower than with the second-generation methods, good quality sequence can be generated by sequencing the same template more than once and on both strand during the run and generating a consensus read sequence from the multiple alignment of all the sequences from each template. The possibility to sequence both strand of the template more than once is provided by the special structure of the library fragments, called SMRTbells. Both end of the DNA fragments are ligated to a hairpin which makes the DNA structurally linear but topologically circular, thus, a strand displacing DNA polymerase will use each position of the circle as template several times. In the case reverse transcriptase is used for sequencing, RNA molecules can be sequenced directly, without the need of cDNA construction. Currently, the PacBio RS II system provides an average read lengths of more than 10 kbp where the longest reads are more than 40 kbp when a library with average fragment length of 20 kb is used.

Nanopore sequencing

Nanopore sensing developed by Oxford Nanopore is an emerging technology which can be used for the detection and identification of a wide range of molecules such as to distinguish the four bases of nucleic acids or to measure protein or ligand concentration (Clarke et al. 2009). The technology relies on the fact that applying voltage across a nanopore seated in an electrically insulating membrane an electric current can be observed. If single nucleotides, strands of DNA or other molecules pass through the pore, the ionic current is disrupted and it can create a characteristic change in the magnitude of the current and enables a direct reading, for example, of the DNA sequence. A nanopore is simply a hole of few nanometers in diameter which can be formed, for example, by proteins like alpha-haemolysin in lipid bilayer membranes or by solid-state materials in synthetic membranes created from graphene or silicon or by a pore-forming protein seated in synthetic material. An Oxford Nanopore device is a sensor array chip which contains tens of thousands of micro wells each of them containing a nanopore and its own electrode. For DNA strand sequencing, DNA is fragmented, then a hairpin is ligated to one end of a fragment and a processive enzyme is bound to the free fragment end. The enzyme associates with the pore forming protein, unzips the double stranded DNA and threads one strand through the nanopore and individual DNA bases on the strand are identified in sequence as the DNA molecule passes through. The presence of the hairpin ensures that the complementary strand is sequenced, too. By using a processive enzyme specific for RNA and adapting the nanopore to distinguish RNA-specific bases, it is possible to analyze the original sample RNA strand directly, without converting it into cDNA.

Application of Next-Generation Sequencing in biomarker discovery

The growth, development and maintenance of an entire organism as well as its responses to the changes of the environment and to biotic and abiotic stresses are directed by its genes, their expression and their interactions. The NGS technology made genome sequencing of any organism affordable and feasible as well as enables genome-wide investigation of variations between individuals, organs, tissues, cells and between different environmental and physiological conditions.

Next-generation sequencing applications include *de novo* or re-sequencing of whole genomes, targeted sequencing of genomic regions of interest, transcriptome and small RNA sequencing as well as genome-wide mapping of DNA methylation or DNA-protein interactions. In addition, the species composition and coding potential of microbial populations can be assessed with the help of NGS.

Genome sequencing

The goal of genome sequencing in the first place is to determine the gene content and to assemble the reference genome of the organism studied then to identify those genetic variations that lead, for example, to better performance, to higher virulence, to resistance to pathogens or to disease susceptibility.

Whole genome sequencing

In *de novo whole genome sequencing* the genome is sequenced and assembled without using direct comparison against a known sequence. After assembling this reference genome sequence, it is possible to perform comparative sequencing or *whole genome re-sequencing* of other individuals of the same species or tissues with genetic abnormalities (such as cancer) to identify polymorphisms, mutations, and structural variations.

A *de novo* genome sequence is always assembled from much shorter reads (50 to 20000 bp) than itself (from kbp to several Gbp). These reads are compared to each other, and the overlapping regions are used to build longer contiguous sequences called contigs. Then scaffolds can be built from neighboring contigs separated by gaps of unknown length (Ekblom and Wolf 2014). Ideally, a genome assembly should deliver end-to-end chromosomal sequences but repeat regions and polyploidy can provide real challenges for whole genome sequencing. Repeated sequences, such as transposable elements or the genes coding for the ribosomal RNA (rDNA) precursor, are distributed throughout the genome. Thus, the shorter reads from these elements can be aligned to many possible regions in the genome and the order of the flanking sequences, *i.e.* the non-repetitive sequences cannot be determined. To overcome this problem and to increase mapping specificity a method called "mate-paired" sequencing is used (Wetzel et al. 2011). Mate-pairs are reads that are originated from the two ends of the same genomic DNA fragment which is typically 1-3 kbp in length but can be as long as 20 kbp. During library preparation adaptors/labels are added to the ends of the fragmented genomic DNA which is then size-selected and the DNA fragments are circularized. The non-circularized DNA fragments are removed by digestion and the circular DNA is fragmented and affinity purified. Finally, the sequencing adaptors are added. For the assembly of most bacterial genomes, current read lengths, accuracy and mate-paired libraries of second-generation technologies are already sufficient to obtain very few contigs (Hunyadkurti et al. 2011). In contrast, the more complex genome of higher eukaryotes, that may contain as long repeats as the members of the Ogre retrotransposon family reaching 25 kb in size (Macas and Neumann 2007), requires other approaches, such as the use of third generation sequencers. Plants in particular provide challenges for genome sequencing. Many plants are polyploid and therefore can have more than two copies of chromosomes. For example, the bread wheat (*Triticum aestivum*) is hexaploid, more precisely, with three very similar, complete diploid genome sets termed A, B and D or the commercial strawberry (*Fragaria x ananassa*) which is octoploid offering real difficulties for genome assembly (Marx 2013). Whole genome assembly of polyploid genomes with only short sequence reads is impossible because it is not known to which chromosome the short read belongs. The combination of third-generation and mate-paired sequencing to create scaffolds and short read sequencing to provide sequencing depth may solve the problem to assemble polyploid genomes.

The reference sequence describes usually the genome of a single individual or in the case of the human genome, of a very few individuals and provides the foundation of genetic studies, however, it cannot explain in itself the phenotypic variations that can be observed in the population. Genetic variations in other individuals on the sequence level can be discovered with re-sequencing of their genome. Usually, high-throughput second-generation techniques providing short reads are used for whole genome re-sequencing to detect single nucleotide polymorphisms (SNPs) and small insertions and deletions (InDels). The employment of third-generation technologies and sequencing of second-generation mate-paired libraries facilitates the recognition of larger insertions and structural rearrangements such as translocations.

One of the greatest achievements of mankind was the sequencing of the reference copy of the human genome 15 years ago (International Human Genome Sequencing 2001; Venter et al. 2001), which created the basis for coping with the big challenge to identify those genomic regions and genes that are responsible for the numerous traits of medical importance, such as driving mutations in cancer, susceptibility to certain diseases or response to drug treatment. There are over 10000 so-called single-gene/monogenic diseases that are known to be determined by the not proper functioning of a single protein resulting from modifications or a single error in the coding gene. Lots of diseases are caused by the complex interactions of more than one gene and have strong genetic component in addition to the environmental influences. Genetic experiments in humans are not allowed by ethical reasons, that is why the identification of the genetic determinants of diseases is attempted and facilitated via re-sequencing based methods such as comparing the genome of normal and tumor cells, investigating family pedigrees and sequencing unaffected, carrier and affected individuals from the families, or performing genome-wide association studies (GWAS) in larger populations.

Since the first studies in which the genomic DNA of tumor and normal cells were sequenced (Ley et al. 2008) several international comprehensive cancer genomics projects (such as the Cancer Genomes Project, the Cancer Genome Atlas and the International Cancer Genome Consortium) have been initiated to identify new driver mutations and to examine the association between mutated genes and the efficiency of drugs. Based on their results more and more targeted therapies to treat patients have become available. GWAS are aimed to discover those genomic regions that are associated with a particular trait or disease and are based on the genotyping of larger populations of several hundred to several thousand individuals with and/or without the investigated trait (such as affected and healthy people) for a huge number (up to millions) of SNPs and structural variants. Variants that are more common or less common, for example, in the affected population may point towards the regions that carry risk or protective variants, respectively. Many structural variants were identified before the NGS era; however, the availability of the massively parallel sequencing technologies provides high-throughput support for the discovery of the structural variants. In 2008, the ambitious 1000 Genomes Project (http://www.1000genomes.org) was launched to create a very detailed catalogue of human genetic variation. The goal of the project is to find the most genetic variants that have frequencies of at least 1% in the populations studied. In the first, pilot phase of the project, whole genome sequencing of 180 samples and 2 mother-father-adult child trios were performed at 2-4-fold and 20-60-fold coverage, respectively (in average, every nucleotide of the genome was sequenced 2-4 or 20-60 times) and the sequence of 1000 gene regions from 900 samples was determined at 50-fold coverage (The-1000-Genomes-Project-Consortium 2010). To complete the sequencing part of the project, 2500 genomes will be sequenced at fourfold coverage.

Targeted genome re-sequencing

There is growing interest in sequencing only specific portions of genomes because targeted re-sequencing strategies provide an efficient and cost-effective solution to analyze these reduced genomes in a highly parallel way. These strategies allow the systematic detection of germline and somatic mutations in cancer or disease associated regions as well as investigation of all type of structural variations having importance in basic or clinical research (Porreca et al. 2007; Choi et al. 2009; Biesecker 2010; Ng et al. 2010a; Ng et al. 2010b). There are a wide variety of target enrichment methods offered for different NGS platforms, to sequence only a limited number of genes or even whole exomes (Gnirke et al. 2009; Bodi et al. 2013; Lin et al. 2014), *i.e.* the transcribed part of all predicted genes. As the exome represents only l-2% of the human genome, even this reduction of the sequencing will significantly decrease the sequencing cost as compared to whole genome sequencing.

Basically, the target enrichment methods are based on two techniques: PCR or hybridization. In the case of amplicon sequencing, which is used to investigate from a few to several hundred genomic regions across multiple samples, multiplex PCR amplifications are carried out with primers specific for the target sequences. An amplicon library can be prepared either by ligating the sequencing adapters to the PCR products or by adding the adapters during PCR amplification by designing PCR primers flanked by adaptors specific for the given sequencing platform to target the genomic regions of interest. The hybridization based methods fall into three classes: DNA-chip-based capture, DNA-probe-based solution hybridization, and RNA-probe-based solution hybridization. To perform the capture of the target sequence, genomic DNA is sheared and processed into a sequencing library specific to a given sequencing instrument. Library fragments are then subject of hybridization to long oligonucleotide probes specific for target regions that are either fixed to a solid surface or are biotin-labelled which can be bound to streptavidin-coated magnetic beads. After the hybridization step, the targeted regions will be immobilized on the solid surface (chip, beads), while the non-target regions remain in the liquid phase and can be washed away. At the end, the library fragments are removed from the solid surface and will be sequenced.

Transcriptome sequencing

Next-generation sequencing is a powerful tool to study the molecular mechanisms and regulation of gene transcription and to catalogue and discover novel transcripts and alternative spliced isoforms (Wang et al. 2009). It gives information about the transcriptional state and gene networks of cells via the quantification and comparison of gene expression levels under different developmental, physiological and environmental conditions. With the transcriptome sequencing approach, different RNA populations including mRNA and small RNAs such as miRNA, PIWI-interacting RNAs, small nucleolar RNAs, small interfering RNAs can be studied and the genome annotation can be facilitated by the determination of the exon sequences. These advantages make RNA sequencing (RNA-Seq) an increasingly attractive method for whole-genome expression studies in many biological systems including species with unsequenced genomes.

Whole Transcriptome Analysis

Whole transcriptome analysis (WTA) via RNA-Seq is aimed to determine the abundance of each coding and non-coding transcripts in samples and enables the detection of genome-wide gene expression changes across different environmental, physiological and developmental conditions as well as in different cell types, tissues, organs of an individual. Because of the high output of the second-generation NGS systems, transcriptome analysis by RNA-Seq is more sensitive, more specific and has higher dynamic range than the microarray based technology. In addition, at proper sequencing depth, it allows the detection of rare transcripts as well as novel and rare splice isoforms (Van Verk et al. 2013).

The quality of starting RNA material highly influences the success of an RNA-Seq experiment. The RNA quality can be measured by Bioanalyser (Agilent Technologies) which gives information about the integrity (RIN) of the RNA sample (ranging from 1-10, from degraded RNA sample to a high quality intact RNA, respectively) (Schroeder et al. 2006). Prior to sequence library preparation ribosomal RNAs (rRNAs) should be removed using a hybridization/bead capture procedure - that selectively removes the rRNA molecules with the help of biotinylated capture probes - from the total RNA sample to enable the detection of less abundant transcripts. The complexity of eukaryotic samples can be further decreased by capturing only those transcripts that have poly-adenyl tail. DNase treatment of the starting RNA material is important to avoid biases coming from genomic DNA contamination. Ribosomal RNA-depleted total RNA is fragmented and converted into cDNA containing the specific adaptors, PCR amplified and sequenced to produce sequence data from one or both end (paired-end) of the cDNA fragment. Finally, the sequencing reads are aligned to a reference genome and counted. Strand specific RNA-Seq protocols permit the detection of antisense transcripts derived from the other strands of the genes. Strand orientation provides information about regulatory relationships that would otherwise be missed and also provides increased confidence in transcript annotation and may increase mapping efficiency.

The third-generation SMRT sequencing technology applies an amplification-free protocol which overcomes the problem of PCR biases and enables a more even coverage of transcripts. Moreover, it sequences an entire RNA molecule from the 5′ to the 3′ end which makes feasible the deep sequencing of full-length RNAs from complex eukaryotic transcriptomes (Sharon et al. 2013).

High-throughput SuperSAGE: Serial Analysis of Gene Expression using NGS

The Serial Analysis of Gene Expression (SAGE) method was developed based on the principles that (i) a short sequence tag of 15-20 nucleotide length can uniquely identify a transcript if the tag is obtained from a unique position of the transcript; (ii) the quantity of a particular tag provides the expression level of the corresponding transcript. As these short tags represent the transcripts, there is no need for WTA by RNA-Seq and in this way, the sequencing costs can be decreased considerably (Matsumura et al. 2010). Polyadenylated RNA molecules or originally non-polyadenylated RNAs, to which a polyadenyl tail was added are captured with biotinylated oligo-dT probes and are reverse transcribed into cDNA. The double-stranded cDNA molecules (still on the surface of the capturing beads) are digested with an anchoring enzyme recognizing four nucleotides (such as NlaIII, DpnII or BfaI) and a platform specific adapter carrying also the recognition sequence of the type III endonuclease EcoP15I of phage P1 next to the anchoring site is ligated to the digested anchored fragments. Then the adapter ligated fragments are digested with EcoP15I cleaving the DNA at a 25 bp distance and with the ligation of a second, barcoded sequencing adaptor and 3-10 PCR cycles, the library is constructed. After sequencing of the library, the reads are mapped one by one on the genome of interest and the transcripts are quantified providing an efficient way to construct differential expression profiles. The drawback of the method is that it is not so robust to detect rare mRNAs and as it targets the 3’ untranslated region of the transcripts fails to discover and survey splice variants.

Small RNA Sequencing

In most organisms, small non-coding RNAs (sncRNAs) play important role in the transcriptional and post-transcriptional regulation of gene expression including the regulation of transposon and foreign DNA activity (*e.g.*, miRNAs, siRNAs, tasiRNAs, piRNAs) as well as in the formation of the spliceosome complex (snRNAs) and in the methylation and pseudo-uridylation of other (such as ribosomal) RNAs (snoRNAs). The regulation of gene expression by the 18-30 nt long sncRNAs is achieved by silencing genes via complementary base pairing with target RNAs (Filipowicz et al. 2008). In this way, most cellular and even organism level processes, like developmental timing, cell differentiation, nutrient signaling and implicated in many diseases is regulated by sncRNAs. Small RNA-Seq allows the identification of already known and novel small RNAs, enables the investigation of their differential expression among different samples by preserving strand specific information. Prior to sequencing small RNA enrichment is recommended. The library preparation workflow is the same as for WTA by RNA-Seq, the only difference is that in the case of small RNAs there is no need for RNA fragmentation.

Metagenomics

Metagenomics is a powerful tool for exploring and comparing complex microbial communities in environmental or clinical samples and enables the study of yet uncharacterized microorganisms which cannot be cultivated in laboratory and represents the vast majority of environmental microbial communities on Earth (Blainey 2013). All organisms including humans live in close association with microbial communities as part of an interdependent metaorganism. The collective genome of the microorganisms that reside in an environmental niche, such as the human gut or skin, is called the microbiome. There are ~10 times more microbial cells (as part of different microbiomes) in (and on) a human body than human cells and the importance and impact of these communities on human physiology and health have been recognized recently. For example, the comparison of the microbiome from lean and obese twins revealed significant differences in the composition, diversity and biochemical potential of the communities (Turnbaugh et al. 2009).

A metagenomics project based on NGS technology can create a catalogue of microbial species that are present in a sample and determine what their ratio is or may survey the biochemical potential of the community in a very cost effective way. In addition, comparative metagenomic analysis of samples collected at different time points provides information about the dynamics of the population. The determination of the identity and abundance of the microorganisms in the sample can be achieved by both 16S ribosomal RNA (16S rRNA) amplicon sequencing and shotgun whole genome sequencing approaches (Sanschagrin and Yergeau 2014). The most commonly used approach for investigating environmental prokaryotic diversity is the amplification and sequencing of the highly conserved phylogenetic marker, the 16S rRNA gene. The 16S rRNA gene contains conservative and variable regions (Pereira et al. 2010) that can be used for designing primers to allow the PCR amplification and after sequencing the identification, respectively. The limitation of the 16S rRNA gene-based sequencing is the primer bias and that it detects the predominant members of the community but may not give information about the rare microbial species because of the low sampling depth. To overcome these limitations, metagenome shotgun sequencing is performed, which targets all DNA that can be extracted from the sample. For this purpose, the extracted genomic DNA is randomly sheared into smaller fragments and a sequencing library specific for the NGS platform used is constructed. Depending on the sequencing depth, not only the composition of the population but also the gene content of the metaorganism can be determined. The challenges and limitations of whole metagenome strategies are the relatively large amounts of starting material needed, the high number of genes with unknown function and annotation and in the case of clinical samples, the potential contamination of metagenomic samples with host genetic material which lowers sequencing depth (Petrosino et al. 2009).

Beside the DNA-based metagenomics, the metatranscriptome of the microbial communities can also be studied. It can provide information about not only the composition of the population but also about the metabolic potential and dynamics of the community, how the microbial activities are regulated and respond to the environmental changes. For metatranscriptome analysis, the environmental RNA samples are converted to cDNA and sequencing platform specific libraries are constructed. For the determination of the composition of the microbial community, the removal of the ribosomal RNA population constituting >95% of the RNA pool is not necessary but for the investigation of actively transcribed genes, rRNA sequences are removed in order to get enough mRNA sequence reads.

Genome-wide mapping of DNA-protein interactions

Protein-DNA interactions play important role in the regulation of gene expression. Chromatin re-modelling is a dynamic process driven by protein modifications that change DNA-protein interactions and alter DNA accessibility for transcriptional regulators (TFs: transcription factors) acting on specific sequences of the DNA. Histone proteins are the primary protein component of chromatin that compact chromosomal DNA. The basic structural and functional unit of chromatin is the nucleosome, which contains four basic proteins - H2A, H2B, H3, and H4 - forming an octamer and about 166 base pairs of DNA that is wrapped around the octamer. Posttranslational covalent modifications of histone tails such as methylation, acetylation, phosphorylation, ubiquitination and sumoylation are important epigenetic regulators of gene expression by determining DNA accessibility for TFs via maintaining an open or closed chromatin state (Campos and Reinberg 2009). Genome-wide mapping of histone modifications and transcription factor-binding sites can be identified by chromatin immunoprecipitation coupled to NGS sequencing (ChIP-Seq) which has become a frequently used technique (Kharchenko et al. 2008; Valouev et al. 2008).

In this technique, the DNA bound to a protein of interest is enriched by immunoprecipitation using specific antibodies raised against the native or antigen-tagged or post-translationally modified protein and the captured DNA is sequenced by NGS. Prior to ChIP-Seq cells or tissues are treated with a cross-linking agent like formaldehyde to covalently attach proteins to DNA. This is followed by cell lysis and fragmentation of the DNA by enzymatic digestion or sonication to approximately 250 bp pieces (a bit longer than the DNA on the nucleosome) and the complexes are immunoprecipitated with the specific antibodies. After immunoprecipitation the cross-links are reversed, the enriched DNA is purified and prepared for high-throughput sequencing by attaching the specific adaptor sequences. The ChIP-DNA is analyzed by NGS sequencing which permits comprehensive coverage and resolution even in the case of large genomes. The technical limitations of the method are the need for a specific antibody raised against the protein of interest and the specificity of the antibody.

Genome-wide mapping of DNA methylation

Epigenetic modifications like DNA methylation play an important role in gene regulation and chromatin remodeling and thereby, are crucial to normal development, differentiation of distinct cell lineages of the organism. Methylation of cytosines is a covalent modification of DNA which usually occurs at the carbon-5 position of cytosine residues (5mC). Methylation of cytosine nucleotides in DNA are known to alter DNA accessibility and thus to repress gene transcription. In mammals, most cytosine methylation occurs at CpG dinucleotides that often in so-called CpG islands that are regions with a high frequency of CpG sites. These CpG islands are often located upstream of promoters or within the gene body (Ehrlich et al. 1982). In plants, DNA methylation commonly occurs at cytosine bases in all sequence contexts: symmetric CG and CHG (in which H = A, T or C) and asymmetric CHH context (Henderson and Jacobsen 2007). Unlike in mammals, DNA methylation in plants predominantly occurs on transposons and other repetitive DNA elements (Zhang et al. 2006). There are two hypotheses how epigenetic modifications can regulate gene expression: i) DNA methylation can form a physical barrier for binding of transcription factors, ii) methyl-CpG-binding domain proteins (MBDs) can bind to methylated DNA and recruit chromatin remodeling proteins like histone deacetylases that can modify histones forming compact, inactive heterochromatin. There are other epigenetic modifications such as cytosine hydroxymethylation (5hmC) and adenine methylation on the carbon-6 position (m6A) but their exact function is not known.

The high-throughput detection of DNA modification has been available only for 5mC by bisulfite sequencing, but nowadays, the third-generation SMRT technology (discussed earlier) makes it possible to recognize other base modifications, too (Flusberg et al. 2010; Fang et al. 2012). Bisulfite sequencing is based on the sodium bisulfite catalyzed chemical deamination of the unmethylated cytosine (C) residues to uracil (U) which is recognized as thymidine (T) by DNA polymerases used in the following steps while 5mC residues are not converted and thus, are recognized as cytosines (Clark et al. 1994). During whole genome bisulfite sequencing (WGBS), either the library (SOLiD) or the DNA before library preparation (Illumina) are treated with bisulfite. After sequencing, reads are aligned to an *in silico* converted, modified reference genome sequence in which the cytosines are exchanged to thymidines and methylated bases are recognized as SNPs. To decrease sequencing costs, enrichment methods to isolate methylated DNA, which are based on DNA-protein interactions, have been developed. Methylated DNA immunoprecipitation sequencing (MeDIP-Seq) uses an antibody that was raised against 5mC residues (Weber et al. 2005; Ruike et al. 2010), while MBD-Seq employs the Methyl-CpG Binding Domain protein to bind methylated DNA (Brinkman et al. 2010; Serre et al. 2010) followed by the immunoprecipitation of the protein-DNA complex with specific anti-MBD antibody-conjugated beads.

Genome-wide mapping of DNA accessibility and chromatin structure

As discussed earlier, the dynamic process of chromatin remodelling plays an important role in transcriptional regulation via changing the accessibility of different DNA regions (and thus, genes) for transcription factors. A number of different methods have been developed to map the position of nucleosome-bound and free DNA.

Micrococcal Nuclease sequencing

Micrococcal nuclease sequencing (MNase-Seq) relies on the characteristics of the enzyme to specifically digest those DNA sequences that are not protected by a protein such as the linker region of the chromatin between nucleosomes (Telford and Stewart 1989). Therefore, the positioning of the nucleosomes and non-histone DNA-binding proteins determines which DNA fragments are protected from digestion and, thus, released from the complex during library preparation and sequenced. The efficiency of MNase digestion depends on the degree of chromatin compaction, *i.e.* less compact chromatin is more easily digested by the enzyme. To perform MNase-Seq, chromatin substrates are prepared from isolated cell nuclei or permeabilized cells and are then treated with MNase which is followed by the purification of the DNA and library construction.

Assay for transposase-accessible chromatin using sequencing

The assay for transposase-accessible chromatin using sequencing (ATAC-Seq) also provides information about chromatin compaction, nucleosome positioning and maps the genomic location of DNA-binding proteins in regulatory regions (Buenrostro et al. 2013). The method is based on the direct *in vitro* transposition of a modified Tn5 transposon carrying sequencing adaptor sequences into the native chromatin. Transposons can integrate only into the active regulatory elements where the chromatin is open. Sequencing reads are generated at locations of nucleosome-free, active regulatory regions and provide base-pair resolution of the open chromatin regions in the genome.

Formaldehyde-assisted isolation of regulatory elements

Formaldehyde-assisted isolation of regulatory elements (FAIRE-Seq) is an alternative method for the genome-wide detection of nucleosome-depleted, open chromatin regions that are associated with regulatory activity, *i.e.* promoters and transcriptional start sites (Giresi et al. 2007). The technique is based on differences in cross-linking efficiencies between nucleosome-bound DNA and nucleosome-depleted regions. In this method, DNA-protein complexes are cross-linked *in vivo* using formaldehyde and then the sample is lysed, sonicated and the DNA is extracted. During phenol/chloroform extraction, the DNA not cross-linked to protein will be recovered in the aqueous phase, while the DNA covalently linked to proteins will become trapped between the organic and aqueous phase. Finally, a sequencing library is constructed from the DNA recovered from the aqueous phase. Sequencing provides information about regions of DNA that are not occupied by histones.

NGS in the clinical practice

NGS technologies have revolutionized our process of learning about the relationship between genetic variations and causes, pathogenesis, progression and prognosis of diseases. While NGS keeps broaden this knowledge and the technologies, methods are continuously improved and developed, sequencing (equipment and operational) costs continue to decline. These developments support the introduction of NGS technology into the clinical practice to facilitate personalized medicine (including bedside diagnosis, decision making on treatment, patient stratification, etc.). Indeed, on 19 November 2013, the U.S. Food and Drug Administration (FDA) approved the use of the Illumina MiSeqDx platform for the diagnosis of cystic fibrosis and certain health insurance plans may cover NGS-based diagnostic investigations. In addition, clinics and diagnostic companies using NGS help decision making in a wide range of clinical environments by supporting activities such as pre-implantation genetic screening, non-invasive prenatal testing, the diagnosis of inherited diseases of single-gene etiology (Mendelian diseases), oncology mutation screening or HLA typing before transplantation.

The success of *in vitro* fertilization (IVF) is most often hindered by chromosome aneuploidy, *i.e.* when the cells contain one or more chromosomes not in two copies. Most embryos with aneuploidy do not implant and those that do often lost by spontaneous abortion. Pre-implantation genetic screening by NGS allows the quick ploidy testing of all somatic and sexual chromosomes to select euploid embryos for implantation and, thus, to increase ongoing pregnancy rate and to enable single embryo transfers which decreases the number of high-risk multiple pregnancies.

Fetal aneuploidy and other chromosomal aberrations affect almost 1% of natural pregnancies, too, and their probability increases with the age of the parents. Classical and even more recent prenatal tests such as karyotyping and chromosome microarrays, respectively, require an invasive process (chorionic villus sampling or amniocentesis) to obtain fetal cells which imposes risks to both the fetus and the mother. The discovery of intact fetal cells in the maternal blood and later the recognition that the fraction of fetal DNA in the maternal plasma cell-free DNA (cfDNA) can be as high as 15% have led to the development of non-invasive prenatal testing (NIPT) methods to detect chromosomal aneuploidies and even subchromosome abnormalities (Fan et al. 2008; Srinivasan et al. 2013). At present, commercially available NIPT services are based on Illumina short-read sequencing of the maternal plasma cfDNA and provide information about the gender of the fetus and about trisomies of chromosomes 21, 18 and 13 leading to Down-, Edwards- and Patau-syndrome, respectively. By the improvement of the analysis tools, however, trisomies with lower probability (such as trisomy 8 causing Warkany syndrome 2) or trisomies (like 16 or 9) often resulting in first-trimester miscarriage as well as chromosomal microdeletion syndromes can be detected, too.

As mentioned earlier, there are over 10000 monogenic diseases that appear with very low to rather high frequencies; cystic fibrosis (CF), for example, affects about one out of every three thousand new-borns among people of Northern European ancestry. Genetic risk factors for certain tumors and for complex traits have also been discovered. Based on the available genetic information, there are a number of NGS based assays, for example, to identify the carriers and to confirm the diagnosis of CF or detect germ-line mutations that predispose individuals to cancer such as the *BRCA* (*Breast Cancer*) gene mutations. There are more comprehensive tests investigating high number of genes that are most appropriate for situations where the evaluation of multiple genetic markers may clarify or refine the diagnosis because the presenting set of symptoms and tests are inconclusive, there are a large number of candidate genes to evaluate, or the phenotype might indicate multiple genetic conditions. In oncology, sequencing oncogenes and tumor suppressors may identify mutations that are targeted with a specific therapy or predispose the therapy ineffective, thus, it helps the decision on the treatment.

Viruses and bacteria can cause serious outbreaks in any population and in even in major clinical centers. The quick reconstitution of how a new strain evolved to cause an outbreak such as the Shiga-toxin producing *E. coli* strain in Germany in 2011 (Mellmann et al. 2011), or tracking the route of an infection, for example, by an antibiotic resistant strain in an NIH hospital (Snitkin et al. 2012) have become possible because of the NGS technology.

Acknowledgements

The publication of this review was supported by the European Union, co-financed by the European Social Fund, within the framework of "Practice-oriented, student-friendly modernization of the biomedical education for strengthening the international competitiveness of the rural Hungarian universities" TÁMOP-4.1.1.C-13/1/KONV-2014-0001 project. Human exome sequencing at the Seqomics Ltd was co-financed by the European Union (FP7 program "Rational molecular Assessments and Innovative Drug Selection"; grant number: 304810) and the Hungarian government (EU_BONUS_12-1 -2012-0034). The authors thank Alina Kereszt for preparing the figures.

References

Biesecker LG (2010) Exome sequencing makes medical genomics a reality. Nat Genet 42:13-14.

Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev 37:407-427.

Bodi K, Perera AG, Adams PS, Bintzler D, Dewar K, Grove DS, Kieleczawa J, Lyons RH, Neubert TA, Noll AC, Singh S, Steen R, Zianni M (2013) Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech 24:73-86.

Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG (2010) Whole-genome DNA methylation profiling using MethylCap-seq. Methods 52:232-236.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10:1213-1218.

Campos EI, Reinberg D (2009) Histones: annotating chromatin. Annu Rev Genet 43:559-599.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096-19101.

Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. Nucleic Acids Res 22:2990-2997.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265-270.

Drossman H, Luckey JA, Kostichka AJ, D'Cunha J, Smith LM (1990) High-speed separations of DNA sequencing reactions by capillary electrophoresis. Anal Chem 62:900-903.

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res 10:2709-2721.

Ekblom R, Wolf JB (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026-1042.

Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 105:16266-16271.

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE (2012) Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. Nat Biotechnol 30:1232-1239.

Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9:102-114.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7:461-465.

Gilbert W, Maxam A (1973) The nucleotide sequence of the lac operator. Proc Natl Acad Sci USA 70:3581-3584.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17:877-885.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182-189.

Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 210:1518-

1525.

Henderson IR, Jacobsen SE (2007) Epigenetic inheritance in plants. Nature 447:418-424.

Hunyadkurti J, Feltoti Z, Horvath B, Nagymihaly M, Voros A, McDowell A, Patrick S, Urban E, Nagy I (2011) Complete genome sequence of *Propionibacterium acnes* type IB strain 6609. J Bacteriol 193:4561-4562.

International Human Genome Sequencing C (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Jarvie T (2005) Next generation sequencing technologies. Drug Discov Today Technol 2:255-260.

Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351-1359.

Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW (2010) Real-time DNA sequencing from single polymerase molecules. Methods Enzymol 472:431-455.

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682-686.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66-72.

Lin MT, Mosier SL, Thiess M, Beierl KF, Debeljak M, Tseng LH, Chen G, Yegnasubramanian S, Ho H, Cope L, Wheelan SJ, Gocke CD, Eshleman JR (2014) Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. Am J Clin Pathol 141:856-866.

Macas J, Neumann P (2007) Ogre elements--a distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390:108-116.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Marx V (2013) Next-generation sequencing: The genome jigsaw. Nature 501:263-268.

Matsumura H, Yoshida K, Luo S, Kimura E, Fujibe T, Albertyn Z, Barrero RA, Kruger DH, Kahl G, Schroth GP, Terauchi R (2010) High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. PLoS One 5:e12010.

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6:e22751.

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J (2010a) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42:790-793.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010b) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30-35.

Pereira F, Carneiro J, Matthiesen R, van Asch B, Pinto N, Gusmao L, Amorim A (2010) Identification of species by multiplex analysis of variable-length sequences. Nucleic Acids Res 38:e203.

Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55:856-866.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J (2007) Multiplex amplification of large sets of human exons. Nat Methods 4:931-936.

Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G (2010) Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics 11:137.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687-695.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441-448.

Sanschagrin S, Yergeau E (2014) Next-generation sequencing of 16S ribosomal RNA gene amplicons. J Vis Exp 90:e51709, Doi: 10.3791/51709.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3.

Serre D, Lee BH, Ting AH (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38:391-399.

Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 31:1009-1014.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674-679.

Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Group NCSP, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 4:148ra116.

Srinivasan A, Bianchi DW, Huang H, Sehnert AJ, Rava RP (2013) Noninvasive detection of fetal subchromosome abnormalities via deep sequencing of maternal plasma. Am J Hum Genet 92:167-176.

Telford DJ, Stewart BW (1989) Micrococcal nuclease: its specificity and use for chromatin analysis. Int J Biochem 21:127-137.

The-1000-Genomes-Project-Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature 457:480-484.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5:829-834.

Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci 18:175-179.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304-1351.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63.

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737-738.

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37:853-862.

Wetzel J, Kingsford C, Pop M (2011) Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. BMC Bioinformatics 12:95.

Wu R, Taylor E (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. J Mol Biol 57:491-511.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126:1189-1201.

Author recommended web-resources

Manufacturers constructed videos explaining the sequencing technologies in detail:
https://www.youtube.com/watch?v=rsJoG-AulNE
https://www.youtube.com/watch?v=WYBzbxIfuKs
https://www.youtube.com/watch?v=womKfikWlxM
https://www.youtube.com/watch?v=nlvyF8bFDwM
https://www.youtube.com/watch?v=NHCJ8PtYCFc
https://www.youtube.com/watch?v=3UHw22hBpAk