Chapter 5
High-Throughput Technologies: DNA and RNA Sequencing Strategies and Potential

Thomas Rio Frio

5.1 The Next-Generation Sequencing and Its Impact on Genomics and Clinical Genetic Testing

5.1.1 Early DNA Sequencing

Nucleic acid sequencing is a key tool for scientific research and clinical diagnosis to understand and decipher the code to all biological life on earth as well as to understand and treat genetic diseases. DNA and RNA are made up of four chemical bases or nucleotides arranged in different ways to constitute the genes. Sequencing is the process to determine the order of these nucleotides in one or more targeted regions. Several sequencing methods were developed in the 1970s, and the approach developed by Frederick Sanger in 1977, rewarded by a Noble Prize in chemistry in 1980, revolutionized the field of genomics (Sanger et al. 1977). The method called chain-termination method or dideoxynucleotide sequencing consists in the polymerization of DNA fragments complementary to a DNA template, using a mix of deoxynucleotide triphosphate (dNTP), the building blocks for DNA, and fluorescent dideoxynucleotide triphosphate (ddNTP). These ddNTPs or terminators emit each light at different wavelengths depending on their nature (ddATP, ddTTP, ddCTP, ddGTP) and prevent the addition of other dNTPs. The synthesis of complementary DNA strand stops when a ddNTP is randomly incorporated, leading to as many different fragments ending each with a fluorescent ddNTPs as the length of the DNA fragment to sequence. Products from this sequencing reaction are then migrated by electrophoresis in capillaries to separate fragments based on their size,

T. Rio Frio, PhD
Genoma SA, NGS Platform, Geneva, Switzerland
e-mail: thomas.rio-frio@genoma.com

C. Le Tourneau, M. Kamal (eds.), *Pan-cancer Integrative Molecular Portrait Towards a New Paradigm in Precision Medicine*,
DOI 10.1007/978-3-319-22189-2_5

and a laser reads the ending fluorescent ddNTP. Successive fluorescent signals allow reconstituting the sequence of the fragment analyzed. DNA fragments of maximum 1–1.2 kb can be accurately sequenced. Different versions of Sanger-based sequencer were released, from one capillary to 96 capillaries. The configuration with the highest throughput per instrument allows 96 samples to be run simultaneously with 24–36 runs per day, which represents a throughput of 1–2 million bases per day. The Sanger-based sequencing method has been extensively used in research and for genetic testing since its development and made possible one of the major scientific achievements, the sequencing of the human genome.

5.1.2 The Human Genome Project

The Human Genome Project (HGP) was a large international collaborative research project started in 1990 and that lasted for 13 years. It involved six countries (France, Germany, Japan, China, UK, and USA) and cost $2.7 billion (https://www.genome.gov/10001772). The goal of this project was to decode the human genome and to create a vast resource of detailed scientific information about the structural organization and function of human DNA. Steps of the project consisted first in the generation of the physical and genetic maps of the human genome, second, in the determination of the complete sequence of the human genome, and third in the identification and mapping of genes. The sequencing of model organisms such as *M. Musculus*, *E. coli*, and *D. melanogaster* as well as conducting functional studies to develop genomic-scale technologies were also part of the project. The first draft of the human genome was published in Nature in 2001 (International Human Genome Sequencing Consortium 2001), and the project was declared officially complete in April 2003 with 99 % of gene-containing part of human sequence finished with 99.99 % accuracy (International Human Genome Sequencing Consortium 2004). The HGP revealed that the size of the genome is 3 billion bases and that it contains around 20,500 genes, but this number may still slightly fluctuate. The human reference genome is freely and publicly available for scientific purposes, and further analyses are still ongoing to treat few refractory problems, such as large repetitive DNA regions.

To achieve this project that required the sequencing of long DNA sections, a new approach using the Sanger-based sequencing method, called hierarchical shotgun sequencing, was developed. Genomic DNA was enzymatically or mechanically broken into 100–200 kilobase pieces and cloned into bacterial hosts using vectors. Clones that reconstituted the human genome with the minimum tiling path were selected, sheared in smaller fragments, and sequenced by Sanger-based method. The complete sequence of the human genome was then possible to reassemble based on partial overlap of generated sequences using specific bioinformatics tools (Waterston et al. 2002).

This major scientific achievement marked the beginning of the post-genomic area. Since the completion of the first human genome and with the availability of a

reference sequence, scientific research aims now to annotate and functionally characterize genes; understand gene-gene interactions, regulation of gene expression, and protein-protein interactions; and apply this knowledge to better understand life and to predict, prevent, and cure diseases. Demand for faster and cheaper sequencing methods has since dramatically increased, and powerful genomic tools became a mandatory condition to achieve these goals. Several academic laboratories as well as biotechnology companies worked on high-throughput sequencing methods based on the shotgun approach and gave rise to next-generation sequencing technology.

5.1.3 The Next-Generation Sequencing Revolution

Similarly to Sanger-based sequencing method, next-generation sequencing (NGS) is the result of multiple developments and optimization of existing method before the launch of the first next-generation DNA sequencer by 454 Life Sciences in 2005 (Margulies et al 2005). The core philosophy behind all NGS methods developed so far is the simultaneous sequencing of million to billions of short DNA molecules in a same reaction that dramatically increases the throughput and reduces cost compared to Sanger-based method. The actual most powerful NGS system, the HiSeq X Ten System from Illumina, can produce up to 6 billion of sequences per 3-day runs and per instrument. It represents 1,000 times the throughput of the most powerful configuration of an instrument running Sanger-based sequencing method. From 2001 to 2007, along with the optimization of Sanger-based associated technology, the cost per megabase (1,000,000 DNA bases) of sequence went from $10 K down to $600 (94 % reduction) (https://www.genome.gov/sequencingcosts/). This cost dropped dramatically from $600 in 2007 to $0.1 (99.98 % reduction) within the next 4 years along with the launch of NGS sequencers and their fast evolution. The sequencing of a whole human genome can now be achieved in 1 day with one sequencer and costs $1,000, whereas it took 13 years for an international collaboration and $2.7B to sequence the first human genome. This huge decrease of sequencing costs leads to an extraordinary boost of scientific research and exponential growth of sequencing data produced. Contrary to Sanger-based sequencing for which the sequencing signal is derived from all molecules in the reaction, NGS technology sequences every DNA fragment enriched by sample preparation individually. A mixture of DNA molecules derived from different genomic targets can be then sequenced in a same sequencing run. Furthermore, a molecular barcode is usually added in all fragments derived from each DNA sample during preparation of sequencing templates. Thus, several genomic targets from several individuals can be sequenced simultaneously. Bioinformatic analysis of sequenced data attributes sequencing reads to every sample based on the analysis of the molecular barcode and to the region on the reference genome to find from which sequence it derived. This characteristic of NGS allows the quantification of variants detected in sequences and not only their detection as with Sanger-based method. Detection of low-level mutations, below 10 %, which is the common admitted detection limit of Sanger

sequencing, is therefore possible with NGS that actually has depending on the sequencer a detection limit of 0.5–1 %. Fine genetic analysis of collected tumoral tissues that are usually contaminated with healthy surrounding tissue is possible.

5.1.4 The Impact of Next-Generation Sequencing

5.1.4.1 Scientific Research

Surfing on the success of the Human Genome Project, first NGS sequencers were designed primarily to scientific research and not intended to a clinical use. NGS has been extensively used for *de novo* sequencing of a broad range of biological organisms, from microbial to human (animals, plants, bacteria, yeasts, etc.), and provides insights into genome, epigenome, and transcriptome that allows the understanding of these organisms. Because of mild high throughput, mostly bacteria and yeast genomes that are relatively small (<1–15 Mb) and their transcriptomes were studied (Farrer et al. 2009). In microbiology, NGS has facilitated the study of the correlation between genotype and phenotype in these widely used genetic models that contribute significantly to the understanding of human genome and genetic diseases. NGS also helps to optimize food safety and process and health management (accurate identification of bacterial infection, helps to manage epidemic diseases). Along with the increase of sequencing throughput and bioinformatics development, larger genomes such as human and mouse genomes (~3.10^9 base pairs both) have been extensively sequenced (1000 Genomes Project Consortium et al. 2010) to improve the reference genome sequence, to study genetic diseases, and to decipher the function of genes, their organization, their regulation, and their evolution. For example, targeted sequencing of exome, the complete coding part of the genome, revealed causal genetic mutations for rare congenital syndromes, intellectual disability, autism, and schizophrenia (Rabbani et al. 2012). NGS has been extensively used in cancer research in the past few years to detect a number of novel cancer-related genes with the sequencing of large sample cohorts through international collaborations. To achieve a high-resolution view of cancer genomes, several NGS-based methods were developed (whole genome sequencing, exome sequencing, transcriptome sequencing, ChIP-seq, etc.),and the combination of these technologies allowed the detection of novel genetic alterations, point mutations, small insertions or deletions, copy number alterations, and structural variations that contribute to oncogenesis, tumor development, and metastasis (Shyr and Liu 2013). NGS was also able to get insight in the tumor genomic intra-heterogeneity through its high detection and quantification resolution. More complex and larger genomes such as in plants (10^9–10^{11} base pairs) often contain large portions of repetitive sequences and transposable elements. Furthermore, polyploidy is common in these organisms, adding another layer of complexity in the study of such genomes. The use of NGS made possible the sequencing of these genomes with multiple genomes released since 2007 (rice, maize, banana, cacao, etc.). Variants influencing significantly valuable

phenotypic traits have been identified, and the agriculture benefits from these findings (Berkman et al. 2012). In archeology and paleontology, NGS has been successfully used for the sequencing of ancient DNA to reconstruct patterns of evolution and to study population genetics and paleontological changes. For example, the generation of high-quality genomes of a Neanderthal individual and other archaic hominins was performed with the use of NGS and helps to clarify temporal and spatial human evolution (Sánchez-Quinto and Lalueza-Fox 2015). NGS has had also a big impact on metagenomics, which is the study of the total genomic content of a microbial community, for example, in the human gut. The high-throughput capability, relatively low cost, and depth of next-generation sequencing make such study easier by avoiding the need to isolate populations who introduced experimental biased. Through NGS approaches, new species were detected and microbial diversity as well as relative abundance could be assessed with a high resolution that has contributed to the understanding of species and functions present in a microbial community (Thomas et al. 2012). A lot of various scientific research fields have benefited from this powerful genomic tool, and further applications are continuously developed to enlarge the spectrum of NGS applications in research.

5.1.4.2 Clinical Genetic Testing

In the years 2000, Sanger-based sequencing was the gold standard for clinical genetic testing because of its high robustness proven through decades of results in research and clinical genetic testing, its ease of use, and ease of analysis. The complexity to run NGS sequencers, the significant failure rate, mild sequencing quality (reduced length of sequencing reads, significant error rate), as well as few available bioinformatics tools prevented the setup of these instruments for clinical diagnosis. The throughput was also too high to simply translate clinical genetic tests performed on Sanger-based sequencers as this time leading to increase costs for similar test. Bioinformatics analysis of generated sequence data was in its early days; no robust ready-to-use and user-friendly solution was available to ensure highest specificity and sensitivity of genetic tests. As for any technological revolution, NGS sequencers evolved quickly with chemistry, workflow, and bioinformatics, with significant optimizations released on a yearly basis. Quality of generated data dramatically improved, opening the possibility of using NGS sequencers in a clinical setting. In parallel to the development of sequencers with higher throughput dedicated to scientific research, sequencing manufacturer developed clinically oriented sequencers with the goal to conquer the highly profitable clinical market. These sequencers were designed as benchtop sequencers. They are relatively small and user-friendly equipment running with reduced manual operations, and they require little maintenance. The throughput is also reduced to fit with genetic testing requirement, the main goal being not the sequencing of genomes but to sequences selected targets with clinical interest. In the meantime, bioinformatics solutions improved significantly and were proven to be robust for the detection of variants. In 2013 the US Food and Drug Administration authorized broad clinical use of the MiSeqDx DNA

sequencing system from Illumina, the first NGS sequencer to be certified for *in vitro diagnostic* use (http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm375742.htm). Later, in 2014, the FDA cleared the Ion Torrent PGM as a medical device (http://news.thermofisher.com/press-release/life-technologies/thermo-fisher-scientific-announces-listing-ion-pgm-dx-system-us-fda-). However, certifications remain limited regarding the versatility of NGS sequencers since only the sequencing of cystic fibrosis-associated gene *CFTR* on the MiSeq is certified and only few run configurations and throughputs are certified. Consequently, many laboratories and biotechnology companies across the world have developed their own sequencing NGS-based tests such as gene sequencing panels and run those in clinical routine despite a lack of certification of NGS processes.

Several companies such as Life Technologies, Illumina, Agilent, and Roche developed reliable and robust solutions to enrich genomic and RNA targets. They and others propose ready-to-use gene enrichment panels with clinical interest or user-friendly solutions to design customized gene panels. For example, the enrichment of human exome that represents the coding part of the genome is used both in research and clinics to search for unsuspected molecular abnormalities in case of genetic disease, and the sequencing of *CFTR* gene is performed for the detection of cystic fibrosis. With these targeted sequencing solutions, laboratories could move from the analysis of a portion of a gene to the screening of a full gene and even more to tens of genes at the same time allowing a most comprehensive clinical screening which was restricted because of cost, time, and hard setup. Sanger-based method is still used but mostly to confirm variants detected by NGS that could be challenging for some NGS sequencers, for example, an insertion/deletion in long homopolymer stretches. Many clinical genetic centers have already reconfigured their diagnostic processes and adopted NGS as the preferred technology for the diagnosis of diseases (genome, exome, or targeted sequencing). For example, Children Hospital of Philadelphia started to run clinical exomes on children with rare conditions, unexplained despite extensive investigations (http://pediseq.research.chop.edu/). Exome has been extensively used in research, and actual knowledge makes it valuable for the assessment of genetic diseases in newborns. This type of study would permit the estimation of the benefit of exome sequencing compared to conventional medical approach as well as its economical impact.

Actually, the main clinical NGS-based application is noninvasive prenatal testing (NIPT) with an estimated one million of samples processed in the world in 2014 and with an expected market value of $3.62 billion in 2019 (Dondorp et al. 2015). In the 1990s, cell-free DNA derived from placental cells has been found to circulate in the maternal blood during pregnancy (Lo et al. 1997). The fetal DNA represents at least 4 % of total cell-free DNA at 10th week of pregnancy subsequently increases along with maternal age and totally disappears in few hours after delivery (Wang et al. 2013; Bianchi et al. 2012, 2014). The tri-test currently performed during the first trimester of pregnancy is based on biochemical analysis and measurement of nuchal translucency. It gives a high portion of false positives and false negatives. False-positive cases are referred to unnecessary amniocentesis with a risk of miscarriage of 1 %, and false-negative cases prevent from the evaluation of a possible termination

of pregnancy. With the high throughput and quantitative property of NGS sequencers, sequencing circulating cell-free DNA extracted from maternal blood plasma and detecting the presence of chromosomal numerical abnormalities with a sensitivity and specificity close to 100 % are possible. NIPT is actually able to call trisomies of chromosomes 13, 18, and 21 as well as sex chromosome aneuploidies. It is also able to determine the sex of the fetus with accuracy close to 100 % (Devers et al. 2013). Briefly, circulating cell-free DNA (mother + fetus) is sequenced and reads counted on each chromosome. After normalization on chromosome length and GC content, a departure from expected quantity of reads corresponding to half of the fetal fraction indicates an aneuploidy of the fetus. NIPT remains informative since the fetal DNA analyzed is derived from the placenta and not from the fetus. Despite the high sensitivity and specificity of NIPT, some false-positive cases happen mostly because of confined placental mosaicism or vanishing twin. An amniocentesis is recommended in case of a positive result, but the low rate of false negative and false positive compared to conventional tri-test makes NIPT a real improvement for prenatal screening. Most recent version of NIPT can detect microdeletions with a resolution down to 5 Mb that is at least equivalent if not better than amniocentesis. However, NIPT as well as other genetic test currently performed using NGS are considered as laboratory developed tests which neither need to be regulated nor certified by FDA in the USA. Major sequencing players have invested this business such as Verinata (Illumina), BGI, Ariosa (Roche), as well as new biotechnology companies such as Genoma and Premaitha. The fast implementation of NIPT in the clinical practice worldwide demonstrates that NGS has started to revolutionize clinical genetic diagnosis. Many other NGS clinical applications are actually evaluated to determine their contribution to diagnosis and treatment selection.

In clinical oncology, whole genome sequencing is considered as the unbiased gold standard to give a high-resolution view of alterations and structure of the genome (Bennett et al. 2014; Bianchi et al. 2014). Because of the size of the human genome and the associated cost, a comprehensive study of the heterogeneity of analyzed tumor remains difficult to perform routinely. Actually, analysis is restricted to the sequencing of some genetic regions or genes in which alterations are known to promote cancer progression. For example, *BRCA1* and *BRCA2*, two genes strongly implicated in hereditary breast and ovarian cancers, are screened in some laboratories only for some well-documented mutations, and some laboratories analyze the complete coding sequence of these genes. A nonnegligible portion of hereditary breast and ovarian cancers are not linked to these two genes, and the screening of other candidate genes such as *TP53*, *PALB2*, *CHEK2,* etc., allows to perform an all-in-one analysis of patients and to avoid a time-consuming iterative strategy gene after gene. For example, Myriad Genetics, the world leader in *BRCA1* and *BRCA2* analysis, performed their sequencing using Sanger-based method. Myriad shifted this test to NGS and proposed now the analysis of several genes by NGS to evaluate predisposition of some other hereditary cancers (colorectal, melanoma, etc.). With the $1,000 human genome, systematic sequencing of the tumoral genome and normal tissue as reference appears to be the next standard medical practice. Such approach would allow the personalization of treatment based on genomic altera-

tions detected avoiding unnecessary and painful inefficient treatments. Unfortunately, the clinical significance of variants detected in whole genome studies remains actually challenging. Extensive sequencing of cohorts of individuals (1000 genome project, 100,000 UK genomes project, etc.), several types of tumors (The Cancer Genome Atlas) are ongoing and would help to refine analysis in a close future and to give better interpretation of sequencing data. Sequencing of circulating cell-free DNA seems a promising approach to indirectly track tumor progression (Lianos et al. 2015). It has been recently shown that DNA from certain types and stages of tumors circulates in the blood of the patient. Through the analysis of the circulating cell-free DNA by NGS, studies have shown that it was possible to detect mutations previously identified in the tumor through a simple liquid biopsy (Lebofsky et al. 2015). It is still too early to use this technique to detect cancer, but this strategy would be of use to monitor tumor treatment by following the evolution of the presence of the mutation in circulating cell-free DNA of patients.

Primarily dedicated to research, NGS has started to unravel its huge potential for clinical applications. Nevertheless, despite its major role in major scientific achievements, NGS is still in development. Some technological issues remain to be fixed to get sequencing tool that fulfills all quality requirements for high-throughput clinical genetic testing. Furthermore, crucial points need to be addressed regarding the use of NGS in clinical diagnostic, incidental findings in case of exome or genome sequencing, genetic profiling for health insurance, and eugenics. It is actually possible to sequence the whole genome of a preimplantation embryo and to select embryos based on nonmedical traits, such as stature, memory, hair and eye color, or athletic ability. NGS is a powerful genomic tool for genomics that has revolutionized scientific research, but strong regulation of its use in clinical settings is required.

5.2 The Next-Generation Sequencers

Laboratory workflow is similar for all actual NGS platforms from the second generation of sequencers. It is divided in 3 phases: (1) preparation of sequencing libraries starting from purified DNA or RNA, (2) library immobilization and clonal amplification, and (3) sequencing. NGS sequencers are able to reconstitute the sequence of several billions of DNA fragments simultaneously but have some restrictions regarding the length of sequenced DNA fragments and the size of sequencing reads. Indeed, DNA fragments cannot exceed 300–500 bp on average and depending on the instrument. The length of sequencing reads is usually restricted to 100–250 nucleotides depending on the sequencing chemistry used and sequencing speed. The first step of NGS process is to generate DNA fragments that have a size and both extremities compatible with the sequencer. The fragmentation is usually performed through either mechanical or enzymatic shearing and the quality assessed by capillary electrophoresis. Other processes such as targeted enrichment by polymerase chain reaction (PCR) are usually designed to generate molecules

with a compatible size. Sequencer-specific ends are added at both extremities of every fragment by ligation of molecular adapters that are used as starting site for sequencing. These modified DNA fragments constitute a sequencing library, and usually, one library corresponds to a single sample. A molecular barcode that is made of 10–15 nucleotides is usually present in one adapter, allowing the sequencing of a mixture of libraries in a same run. The sequencing of several samples simultaneously could be then easily and safely performed, which represents a significant decrease of sequencing cost and increase of sample throughput. Despite their advanced technology, actual NGS sequencers are not sensitive enough to detect a chemical signal that would be emitted during the sequencing of a single molecule. Every molecule is amplified in close vicinity to produce a localized, uniform, and high signal intensity during sequencing. Thus, libraries are immobilized on a solid support, and clonal amplification is performed through proprietary technologies detailed hereafter. Clonally amplified and immobilized library molecules are then sequenced using sequencer-specific chemistry and strategies.

5.2.1 Illumina Sequencers

Illumina is the actual leader in the development and manufacturing of high-throughput sequencing systems. Most of public and private laboratories own one or more Illumina NGS sequencers. In 2013, Illumina sequencers represented 71 % of all installed NGS sequencers in the world. They are well known for their high-throughput sequencing data, their low error rate, and their reliability even if a lack of diversity in sequenced libraries would decrease significantly the throughput. They are actually considered as the gold standard of NGS. Illumina is an American company funded in 1998 and based in San Diego. After the successful commercialization of a bead array platform for SNP genotyping, gene expression, and protein analysis, Illumina acquired Solexa that developed the sequencing-by-synthesis technology, the technology used by all Illumina sequencers. Prepared sequencing libraries are flowed on a flow cell and are randomly immobilized through the annealing of library adapters to flow cell-coated complementary DNA fragments. Clonal amplification of each DNA molecule of the libraries is performed by bridge amplification PCR to generate isolated clusters of around one million identical single strand fragment. A sequencing primer complementary to the unbound library adapter is hybridized on almost every molecule of every cluster to start the sequencing reaction. The sequencing-by-synthesis technology uses fluorescent-labeled dNTPs that contain a terminator, which prevents the addition of several nucleotides. In every sequencing cycle, nucleotides and polymerase are flowed over the flow cell and one single dNTP is incorporated by the polymerase to each growing strand. Then, a highly sensitive camera scans the entire flow cell to detect the specific fluorescence of the dNTP added at each cluster position. The terminator is then enzymatically removed and another sequencing cycle starts. The sequence derived from every cluster that corresponds to one library molecule is then reconstituted.

The three main sequencers produced by Illumina are the HiSeqs, the MiSeq, and the NextSeq500 (Fig. 5.1) (www.illumina.com). Their price is high ranging from $750,000 for the HiSeqs to $125,000 for a MiSeq (www.allseq.com). HiSeqs are dedicated to large sequencing project such as human exome, whole genome, and transcriptome. Several HiSeq models exist. The widely used HiSeq2500 can generate up to 4 billion single reads of 125 nucleotides or paired-end reads (both extremities of library fragments sequenced) in only 5 days. It represents up to 1 terabase of sequence, enough to sequence 6 human genomes simultaneously. The HiSeq2500 can also be run in fast mode for fast turnaround sample sequencing for clinical diagnosis. It then delivers up to 1.2 billion single reads or paired-end reads of 150 nucleotides in 60 h, which represents 250–300 gigabases of sequence. Two whole human genomes can be sequenced in 60 h using this sequencer. Latest versions of HiSeq released in 2015 have an improved clustering of libraries allowing faster sequencing turnaround time and the increase of the throughput and longer reads. For example, the HiSeq4000 (2 flow cells run simultaneously) is able to generate up to 2.5 billions of up to 150 nucleotide-long reads in 3.5 days, which represents up to 750 gigabases of sequence per flow cell. The HiSeq4000 is therefore mainly dedicated to research projects, sequencing of human genomes, exomes, transcriptome, etc. To sequence human genomes with a very high throughput, two other versions of HiSeq have been released in 2014 and 2015. The HiSeq X five and the HiSeq X ten being

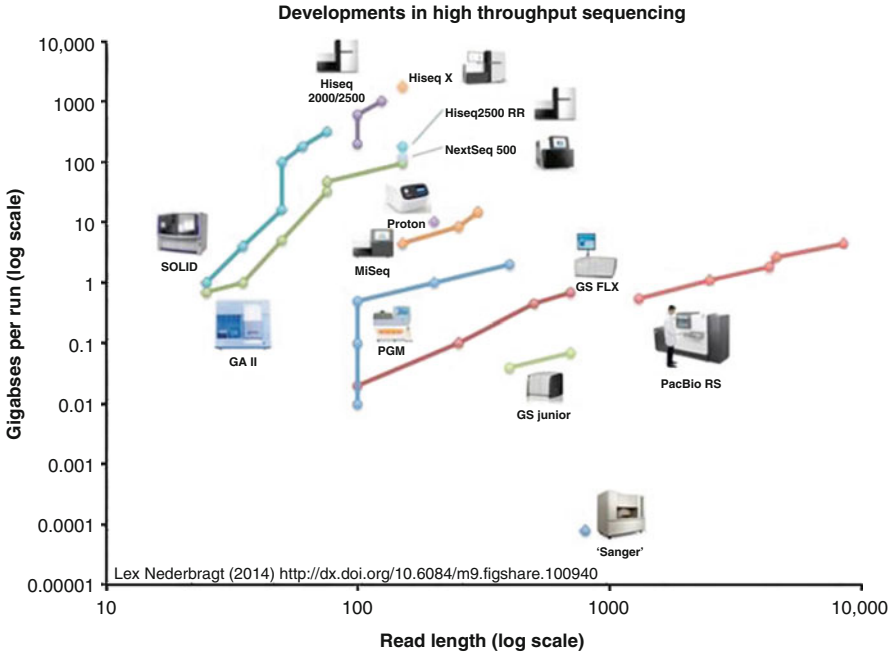

Fig. 5.1 Overview of NGS sequencers. Full run throughput in gigabases (billion bases) is plotted against single-end read length for the different sequencing platforms, both on a log scale

5 and 10 optimized and interconnected HiSeqs. The HiSeq X ten is powerful enough to sequence 18,000 human genomes in 1 year for $1,000 each, inclusive of instrument depreciation, DNA extraction, library preparation, and estimated labor for a typical high-throughput genomics laboratory. Illumina released two benchtop sequencers, the MiSeq and NextSeq500 sequencers. They are dedicated to fast sequencing and long reads with low to middle scalable throughput to match cost to sequencing data amount required. They fit perfectly the needs of the emerging sequencing-based diagnostics market. The MiSeq delivers up to 50 million single reads or paired-end reads in 55 h and with reads up to 250 nucleotides. It is mostly dedicated to targeted resequencing for clinical applications as well as for sequencing of small genomes (bacteria, yeasts) in research. The NextSeq500 delivers up to 800 million reads or paired-end reads of up to 150 nucleotides in 26 h. It is multipurpose sequencer that can sequence human exome or genome as well as small gene panels, ChIP-seq, and other mid-throughput applications.

5.2.1.1 Ion Torrent Sequencers

In 2013, Thermo Fisher Scientific bought the leader in Sanger-based sequencers Life Technologies. Life Technologies released in 2006 their first NGS sequencer, the SOLiD (Sequencing by Oligonucleotide Ligation and Detection) system and the SOLiD 5500 in 2011. Despite significant optimizations of the 5500 system, a high accuracy due to dual reading, the technology failed to move from mid high-throughput to high-throughput as Illumina successfully did. SOLiD sequencers suffered the comparison with HiSeqs because of laborious hands-on, low reliability and throughput, high cost per megabase, and lack of bioinformatics development. Life Technologies acquired then Ion Torrent Systems Inc. in 2010. Ion Torrent Systems Inc. developed an ion semiconductor sequencing technology, a method of DNA sequencing based on the detection of protons released during DNA polymerization. Development of SOLiD systems became since very limited and currently only Ion Torrent systems are sold and further developed by Thermo Fisher Scientific. Ion Torrent sequencers represented 16 % of NGS sequencers installed in the world in 2013, one fourth of Illumina sequencers. Contrary to Illumina, no fluorescence is measured during sequencing reaction but pH. First, libraries are immobilized on hydrogel beads by emulsion PCR. Briefly a single tube PCR reaction containing million of droplets each constituting a single PCR system is performed. In each droplet, one library DNA fragment and one bead are present. The library fragment hybridizes on the bead through the annealing of library adapter to complementary DNA fragments coated on. The PCR replicates more than thousand times the library molecule on the bead. Once the reaction is over, emulsion is broken and beads containing fragments are specifically recovered. This emulsion PCR-based system needs to be improved since more than one library molecule per emulsified PCR system is present in 20–30 % of droplets, leading to a loss of 20–30 % of throughput. Prepared beads are then loaded on Ion chips that are the size of a microprocessor and which contain several millions of wells; each can receive only one bead due

to space limitation. This process has been recently shifted from manual to automatic with the commercialization of the Ion Chef in 2014. The ready-to-sequence chip is then loaded in an Ion Torrent sequencer, either the Ion PGM or the Ion Proton. To sequence the library molecules immobilized on beads, the sequencer flows the 4 unmodified dNTPs, dATP, dTTP, dCTP, and dGTP successively including a wash between every flow. When the flowed dNTP is complementary to the next unpaired nucleotide on the template strand, it is incorporated into the growing complementary strand by the DNA polymerase. A proton is released during this DNA polymerization, and the pH of the well in which the bead stands is modified. Each well of the chip has a corresponding ion-sensitive field-effect transistor that measures ion concentration in solution. The sequencer detects and records the pH modification every time a nucleotide is incorporated in a well. At the end of the run, the recorded signals in every well are transformed into DNA sequence corresponding to library fragments.

The two sequencers actually available, the Ion PGM and the Ion Proton, differ only by their throughput (Fig. 5.1) (www.iontorrent.com). The number of wells present in Ion chips determines the throughput of the sequencer. Three chip formats exist for Ion PGM, 314, 316, and 318 chips that can deliver up to 0.6, 3, and 5.5 million reads of maximum 400 nucleotides in 2–7 h. The Ion Proton has only one chip available, the PI chip, the PII being planned to be released in 2015. PI chip delivers up to 82 million reads of up to 200 nucleotides in 2–4 h. These throughputs are insufficient to sequence large genomes or transcriptomes, but these two benchtop sequencers were designed for the clinical diagnosis market. Their major strengths are the cost of devices ($50,000 for a PGM and $149,000 for a Proton) (www.allseq.com), the rapid sequencing that happens in less than 1 day, a scalable throughput, and low operating prices, in part because of the absence of fluorescence. The main limitations of the system are the absence of high-throughput system and the sequencing of homopolymer regions. Contrary to Illumina sequences for which only one base can be added before signal acquisition, if the same base is repeated on a template strand, then multiple nucleotides are incorporated with the Ion Torrent technology. It leads to the release of a higher rate of protons modifying proportionally the pH. A homopolymer of two consecutive identical bases has a signal twice a single nucleotide (100 % increase) which is easy to quantify, but the difference between signals corresponding to 8 and 9 identical successive nucleotides does not differ enough (theoretical increase of 12.5 %) to avoid miscalling. A lot of work is done by Ion Torrent to improve the accuracy of homopolymer sequencing, notably with the recent release of the Hi-Q chemistry.

5.2.1.2 Roche Sequencers

Roche was acquired in 2007 454 Life Sciences, a company founded by Jonathan Rothberg, the founder of Ion Torrent. The high-throughput sequencing technology developed by 454 Life Sciences is based on pyrosequencing technology and is very similar to Ion Torrent technology, but instead of pH modification, fluorescence is

emitted upon nucleotide incorporation during DNA polymerization. Libraries are prepared similarly than with Ion Torrent method, and prepared beads are loaded on a Pico TiterPlate, a fiber-optic chip, one bead per well. A cocktail of enzymes, DNA polymerase, ATP sulfurylase, luciferase, and apyrase is added in every well as well as their substrates, adenosine 5' phosphosulfate (APS) and luciferin. Similarly to Ion Torrent sequencing devices, the 4 nucleotides are sequentially flowed by the sequencer into the chip, and their incorporation generates a signal recorded by the sequencer. When a nucleotide is incorporated to the growing complementary strand of a library molecule by the DNA polymerase, a pyrophosphate is released in the well. The ATP sulfurylase presents in the well converts this pyrophosphate in ATP in the presence of adenosine 5' phosphosulfate. Through the action of luciferase, luciferin is converted in oxyluciferin that generates an amount of fluorescence proportional to the amount of ATP that corresponds to the number of nucleotides incorporated. Unincorporated nucleotides as well as ATP are then degraded by the apyrase, and another nucleotide is flowed into the chip. At the end of the run, nucleotide sequence of library molecules present in every well is reconstituted. The first 454 sequencer released in 2005 was the Genome Sequencer FLX, and a newer version was released in 2008, the GS FLX+Titanium system (Fig. 5.1) (www.454.com). The last system version generates up to one million reads of up to 1000 bases in 23-h runs for a throughput of 700 megabases (www.allseq.com). This throughput is significantly lower than other sequencers on the market, but the long reads produced make this platform extremely useful for niche applications such as the assembly of *de novo* sequenced genomes. Later on a less powerful version of the Genome Sequencer FLX system, the GS Junior was released. This device can sequence up to 100,000 reads of up to 700 bases in 10 h and is mainly dedicated to researchers with modest sequencing needs. Due to a high cost per base, a high error rate in homopolymers, and low throughput compared to Illumina and Ion Torrent sequencers, Roche announced in 2013 the shutting down of their sequencing business.

5.2.1.3 The Third Generation of Sequencers

The actual second generation of sequencers suffers from bias and limitations mainly due to the requirement of amplification of library molecules before sequencing. Indeed, the signal (fluorescence, pH) detected by the sensor systems of sequencers needs to be intense to be detected. Since nucleotide incorporation does not happen correctly in all molecules of the same cluster or beads, dephasing of sequencing signal occurs along with the growing of sequenced strand and prevents the sequencing of accurate longer reads. Short reads produced by actual sequencers are not sufficient to generate long continuous assemblies of complex genomes that contain numerous repetitive sequences (transposable elements, high copy genes, centromeric/telomeric sequences, segmental duplications). The third-generation sequencers aim to be able to sequence single molecules allowing direct sequencing of nucleic acids, long reads, no bias due to amplification (GC content), and absolute quantification. Currently, only one third-generation sequencer has been

released yet, the PacBio RS by Pacific Biosciences in 2010 and, its latest version, the PacBio RS II in 2013 (Fig. 5.1) (www.pacificbiosciences.com). It can generate reads of up to 15,000 bases in real time but with a reduced throughput of 50,000 reads (up to 1 gigabase sequenced) in up to 240-min run and with a much lower quality compared to second-generation sequencers. Latest version of reagents, protocol of library preparation, and system produce reads with an average length >10 kilobases (www.allseq.com). The optical system that records the sequencing signal is essentially taking a movie of fluorescent nucleotide incorporation. Briefly, single molecule is bound to a single DNA polymerase coated in a zero-mode waveguide (ZMW) on a sequencing small plastic cell called single-molecule real-time cell. ZMW is a structure that captures signal only from nucleotides that are being incorporated, while signal emitted by unincorporated is filtered out. The main applications of this system are for applications that required long reads such as de novo sequencing of small genomes. The rate of nucleotide incorporation is 2–3 bases per second, and the measure of nucleotide incorporation rate allows the determination of modification status of the template nucleotide (5-mC, 5-hmC, etc.), making this sequencer interesting for epigenetic studies. Advantages are low cost of run and single-molecule sequencing, but the main weaknesses are a high machine cost, a low throughput, and low raw accuracy of reads even if contrary to second generation of sequencers; sequencing errors are stochastic and the use of multiple reads gives high accurate consensus reads.

One of the most promising types of third-generation sequencers is based on nanopores. Several companies such as Illumina and Roche are developing or have interest in nanopore-based sequencers. Actually, the most advanced project is conducted by Oxford Nanopore Technologies, a UK-based company that has worked on nanopores for almost 20 years. In 2013, they selected genomic centers to evaluate the technology of their first nanopore-based sequencer, the MinION, which is the size of a USB key (www.nanoporetech.com). It contains biological pores through which DNA molecules pass. It is able to identify bases of DNA by measuring the changes they generate in electrical conductivity when the DNA strands flow through the pore. Sample preparation protocol includes the incorporation of a hairpin adapter that links the 2 strands of DNA molecule by one end. Both strands of a DNA molecule can be sequenced sequentially to generate a highly accurate consensus sequence. After numerous improvements of flow cells and sample preparation kits in 2014, latest released data showed that the MinION could deliver reads with a length up to 150 kilobases with an average of ~5 kilobases (Madoui et al. 2015). Some runs have produced up to 490 megabases of sequence in 48 h. The accuracy remains poor with an average identity (how closely the read matches a reference) of 75–85 % (Madoui et al. 2015). Nanopores are more than a single base in height so that the ionic signal measurements are not of individual nucleotides but of approximately 5 nucleotides at a time. Therefore, the base calling must individually recognize at least $4^5 = 1024$ possible states of ionic current for each possible 5 mer, increasing dramatically the complexity of the signal. Two other nanopore-based sequencers are currently in development by Oxford Nanopore Technologies with

increase throughput, the GridION, and the PromethION which are planned to generate 1 gigabase of sequence per minute.

Several other third-generation sequencers are currently in development, notably the GnuBIO system (Bio-Rad), NabSys sequencer, GeneReader (Qiagen), etc. Some of these systems should revolutionize sequencing as NGS did and consequently genomic scientific research as well as clinical genetic testing with very fast and cheap and reliable sequencing of long DNA pieces.

5.2.2 NGS Applications

5.2.2.1 Genomics

Recent progress in technology led to substantial cost reduction and increased throughput and accuracy of DNA sequencing. A flow of genetic data has continuously grown, and scientists across many fields have used NGS for a multitude of applications (Fig. 5.2). In genomics, sequencing and resequencing of full genomes require a lot of sequencing data but few preparation steps. DNA is extracted and sheared through mechanical or enzymatic action. The library preparation consists in end repair and adapter ligation. A human genome requires at least 100 gigabases of sequences, and smaller genomes such as Escherichia coli require as little as 125 megabases that represents a tiny fraction of the NGS throughput. Sequencing a whole genome is not a standard approach even today for research or clinical applications because of its associated cost despite a huge decrease over the last 7 years. For example, tumor samples are heterogeneous, and standard genome sequencing

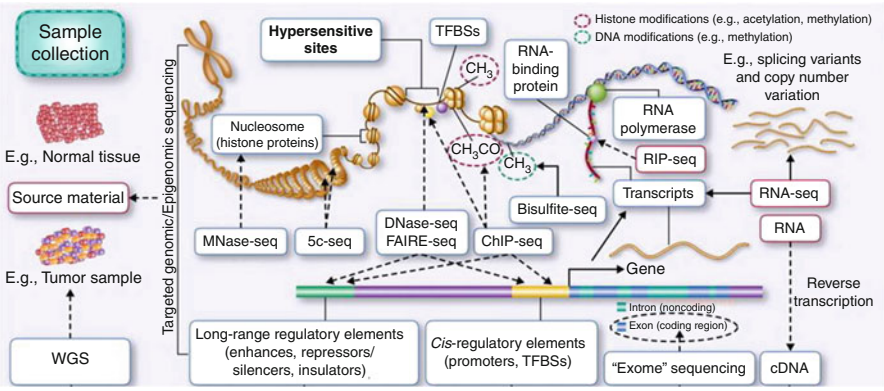

Fig. 5.2 NGS applications. WGS: whole genome sequencing; Mnase-seq: sequencing of nucleosome-associated DNA; 5c-seq (also 4C-seq, HiC-seq): chromosome conformation capture, identification of genome-wide regulatory interactions of a given locus with its unknown partners; DNAse-seq, FAIRE-seq: identification of open chromatin regions; ChIP-seq: identification of protein binding sites; Bisulfite-seq: identification of methylated regions, RIP-seq: identification of protein-RNA binding sites

used for organism genomes does not produce enough data to have a clear picture of tumor-associated molecular events. The depth of coverage which represents the number of time any targeted base is sequenced by independent sequencing reads needs to be around 100×, whereas constitutive genomes are commonly sequenced with a depth of 30×. Therefore, comprehensive sequencing of a tumor genome would cost at least 3 times more.

An alternative is the sequencing of the exome, which represents 1 % of the genome and encompasses all coding regions, or the sequencing of several genes, one gene, or a part of gene. Such targeted sequencing is achieved through the enrichment of the region of interest during library preparation before sequencing. This strategy has a reduced cost compared to the genome and allows deeper investigation of the region of interest even if the targeting strategy could be expensive depending on the method. Enrichment of targeted regions is mainly performed following two different approaches, capture and PCR amplification (Mamanova et al. 2010). The capture method is mostly used for large targets such as the exome or more than 10 genes. After library preparation, library molecules are incubated with probes designed to hybridize with targeted genomic regions. After incubation, DNA-probe complexes are recovered usually by using streptavidin-coated beads that selectively bind the biotin linked to the probes. Thus, the sequenced libraries contain only the targets. A large cumulative size of targets can be sequenced through this method, but the capture of small targets often leads to a significant portion of off-target sequencing data. Main challenges of this method are the DNA's high quality which could be challenging for some samples, long library preparation compared to other NGS applications, and specificity of the capture. Indeed, some thermodynamic constraints prevent from an efficient capture of some genomic regions (high or low GC content, repetitive regions) that lead to poor sequencing of these targets.

Isolation of regions of interest by PCR is usually the preferred method for small size cumulative target length. DNA is mixed with primers that are complimentary to regions of interest. Simultaneous amplification of all regions is performed with the multiplexing of all PCR reactions in one or more reaction tubes. Then, starting from amplified fragments, a library is prepared and sequenced. The main advantages of this technique are the ease of sample preparation and the low amount of DNA required. Furthermore, bad quality samples such as FFPE samples can be successfully processed by reducing the size of PCR products and increasing the number of primers. The main challenge of this approach is the uniformity of sequencing. Indeed, usually hundreds of PCR reactions simultaneously occur, depending on the size of the target, and since efficiency of each PCR reaction is usually nonequal, some targeted regions are poorly sequenced and some highly sequenced. This amplification heterogeneity amplifies along with the number of targets. The under- and over-sequencing has a huge impact on the final cost since the depth of sequencing wanted must be based on the poorest sequenced region. Similarly to capture method some challenging regions could neither be amplified nor sequenced. Other methods for targeted sequencing exist such as the Haloplex approach. It consists in the capture and amplification of targeted regions of DNA sheared through a constant pattern with specific cocktail of enzymes and using existing probes selected in a

catalog. To help customers, exome and some predesigned gene panels have been optimized by manufacturers and can be purchased directly. Custom designs are set up through user-friendly web interface directly on manufacturer websites such as Agilent (SureSelect, Haloplex), Illumina (TruSeq, Nextera custom), and Thermo Fisher (AmpliSeq).

In addition to sequence one or more targeted genomic regions and through its quantitative properties, NGS could be used to assess copy number variants, rearranged tumoral genomes, as well as any structural and numerical chromosomal abnormalities. Data have shown that using sequencing data from exome or even from fewer targets as well as a set of control samples, a complete view of copy number variants in every targeted region for a sample can be produced (Krumm et al. 2012). Hence, NGS avoids the use of complementary experiments to get a comprehensive view of the genome studied, reducing cost and sample use. Another strategy called mate-pair sequencing allows the full study of rearrangements such as translocation using few sequencing power and therefore reduced cost (Korbel et al. 2007). It relies on the sequencing of both extremities of long DNA fragments (usually 5–15 kilobases). Through the comparison of the expected and experimentally determined distances between the two sequenced extremities, genomic structural rearrangements such as large insertion/deletions and translocations can be identified. The fragment length produced for mate-pair sequencing experiment can be experimentally tuned to change the resolution of the analysis and the final cost of the experiment.

5.2.2.2 Transcriptomics

Formerly, RNA expression was performed by quantitative PCR and then by microarray techniques. Quantitative PCR is limited since it cannot be used for whole transcriptome analysis, and microarrays suffer from reduced dynamic range. Next-generation sequencing offers rapid high-throughput gene expression profiling and any RNA type (mRNA, ncRNA, miRNa, piRNA, snRNA, etc.). Despite microarrays still being considered by many as the gold standard for genome-wide expression study, recent data have shown that NGS has a better sensitivity and has started to become the new reference tool for RNA studies since both expression quantification and sequencing can be achieved within a single experiment (Wang et al. 2009; Ledford 2008).

Depending on RNA type studied, library preparation varies. For the sequencing of mRNA transcripts, RNA should be of high quality and not fragmented. First, either mRNA is enriched using beads coated with polydT primers or a depletion of ribosomal RNA is performed. Through the use of polydT primer, only mRNAs are enriched, whereas using ribodepletion, any long RNA molecule will be sequenced, which could increase the experimental noise. PolydT-mediated enrichment is only recommended with high-quality samples since a low representation of 5' end of RNA molecule could be observed with fragmented RNA samples. RNA is then fragmented, a reverse transcription is performed to produce double-stranded cDNA

fragments, and adapters are ligated at both ends. More than 100 million pairs of reads are recommended for a full human transcriptome to study expression, sequence of expressed transcripts, isoforms, as well as rearrangements (fusion, translocation). A study restricted to gene expression quantification requires as few as five million of single reads. RNA studies are usually scaled depending on the resolution needed. To sequence small RNA molecules such as miRNA, long adapters are linked directly onto the miRNA at the beginning to reach a length compatible with library preparation specifications.

The main challenge of RNA sequencing by NGS is that contrary to genomic sequencing, no standards for data control exist. Depending on sample type, the set of genes expressed varies as well as their expression. It is therefore almost impossible to control the quality of results. Synthetic RNAs could be added to the RNA sample before library preparation to evaluate biases linked to library preparation and the resolution of sequenced data. Another actual challenge of RNA sequencing is the analysis of alternative transcripts. Production of small reads by NGS sequencers complicates the characterization of all alternative transcripts even if paired-end reads have improved such analysis. Third generation of sequencers would ease the analysis of alternative transcripts through the production of long reads.

Recently, targeted RNA sequencing was developed to specifically sequence some transcripts or part of transcripts. After reverse transcription of RNA sample, a PCR is performed to enrich specifically for targeted regions. Similarly, detection of gene fusion using pre-designed mix of PCR primer can be achieved.

5.2.2.3 Other Applications

NGS has been found to be a very powerful tool to study protein-nucleic acid interactions. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the most widely use procedure to detect the interaction between proteins and DNA (Park 2009). This technique allows researchers to identify across a whole genome binding sites of a protein of interest that can be transcription factors, DNA-binding enzymes, histones, chaperones, or nucleosomes. Chromatin-bound proteins are cross-linked and the chromatin is sheared. Chromatin fragments linked to the protein of interest are then immunoprecipitated through the use of a specific antibody. After removing nonbound chromatin, de-cross-linking step removes protein to DNA fragments that undergo library preparation and sequencing. Binding sites of the protein of interest are detected by mapping of the sequencing reads on the reference genome. Similarly, RIP-seq (RNA immunoprecipitation) is used to study RNA-protein interaction through a similar process that includes a reverse transcription to convert recovered RNA fragments in cDNA before library preparation. Compared to ChIP-chip assays (ChIP followed by microarray analysis), NGS has a better resolution, low noise, and high genomic coverage.

ChIP-seq is also used for the analysis of histone modifications that play a key role in transcriptional regulation. Active and inactive transcriptional regions of the chromatine, open and compacted chromatin states, that are regulated by specific modifications of histone (methylation, acetylation) targeted by specific

antibody could be identified. To analyze transcriptional regulation, several methods exist such as DNAse-seq (Song and Crawford 2010), FAIRE-seq (Hesselberth et al. 2009), and MNase-seq (Barski et al. 2007). DNAse-seq (DNase I hypersensitive site sequencing) sequences genomic regions hypersensitive to DNAse I that are not packed, therefore implicated in transcription regulation (promoters, enhancers, cis-regulatory elements, etc.). Briefly, chromatin is digested by DNAse I that cuts DNA in non-condensed chromatin regions. Cleaved DNA fragments are purified and, after library preparation, sequenced. Mapping of sequencing reads on reference genome allows the identification of regulatory regions. An alternative to DNAse-seq is the FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements). Formaldehyde is used on chromatin to link proteins to DNA. Chromatin is sheared by sonication and purified using phenol-chloroform solution. DNA that is not linked to proteins is in the aqueous phase, whereas linked DNA is in the organic phase. DNA from aqueous phase is then recovered and sequenced, thus allowing the mapping of regulatory regions across the genome. MNase-seq (micrococcal nuclease) uses the micrococcal nuclease which digests open chromatin regions to enrich for nucleosome-associated DNA (packed regions of chromatin). Sequencing of nondigested fragments reflects protection from MNase and to transcriptionally inactive genomic regions.

Another aspect of epigenetics is the study of the methylation state of DNA. Methylation occurs at CG sites through the addition of a methyl group to the cytosine by DNA methyltransferase. Methylation occurs usually in promoter regions, and the more methylated a promoter region is, the more the expression of the gene is repressed. Methylation pattern is important for embryonic development and cell differentiation. Genome-wide analysis of methylation can be achieved by bisulfite sequencing (Krueger et al. 2012) or meDIP-seq (Ruike et al. 2010), for example. Bisulfite sequencing uses the property of bisulfite to convert non-methylated cytosines into uracil while methylated cytosines are not affected. Sequencing by NGS of whole bisulfite-converted genome is challenging since the low sequence diversity complicates the mapping of reads on the reference genome since the vast majority of cytosines are sequenced as thymines after bisulfite treatment. Bisulfite is the common method to generate a clear picture of the methylome, but it is expensive, about 1.5× the cost of a genome. To lower costs, MeDIP-seq (methylated DNa immunoprecipitation) can be used as an alternative. Methylated DNA is immunoprecipitated with an antibody specific to 5'-methylcytosine. This enrichment of methylated DNA fragments does not alter genomic sequence but introduced some biases. It significantly reduces the sequencing throughput required as well as experimental cost while offering genome-wide coverage of methylation.

5.3 Challenges for the Future

Next-generation sequencing has revolutionized the field of genomics for the last 10 years. Sequencing of whole genomes or transcriptomes, large sample sets in short turnaround time, and reasonable cost have had a huge impact on scientific

research and clinical genetic testing. Nevertheless, the actual second generation of sequencers suffers from limitations and biases that need to be fixed in order to get at least one gold standard technology. The huge reduction in cost has spawned an increasing demand of sequencing so that now scientist can expand sequencing targets with the ultimate goal to sequence only genomes. Targeted sequencing suffers from bias and incomplete coverage of targets as soon as a significant cumulative size of target reached a certain level. Many improvements happened with PCR-free preparation of libraries or low amount of starting material, but these types of library preparations are expensive and introduced non-PCR-based biases. The genome appears to be the next gold standard for genomics research as well as for clinical genetic testing to get a complete picture of someone's genetic background with a minimum of experimental bias. A significant step in this direction has been recently made with the $1000 human genome cost, but another challenge has appeared: the bioinformatics capacity to treat and store such amount of sequencing data. A human genome takes roughly 1 day to be analyzed. A huge increase of genome sequencing would require a strong improvement of bioinformatics softwares to reduce calculus time and storage. It would not be surprising that costs linked to bioinformatics and storage would become similar or even higher than the cost of the sample preparation and sequencing.

The emerging third generation of sequencers shows promising early performances and gives insight to the close future of sequencing. Long reads of ten to hundreds of kilobases would help the assembly of genomes and to get a clear picture of transcriptome. Long sequencing would also improve the detection of large deletion, insertion, and chromosomal rearrangements that are of great importance for diagnostics, for example, in oncology, but still remain challenging today. Reduced error rate will also help scientific research and clinical genetic testing to avoid cross-validation and increase specificity and sensitivity of tests. Detection of variants with low frequency levels (<0.5 %) that have an interest in oncology, for example, subclonal tumoral events that could generate treatment-resistant tumor relapses, would be improved. Similarly, tracking circulating tumoral DNA would benefit patients.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D et al (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073. doi:10.1038/nature09534

Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837

Bennett NC, Farah CS (2014) Next-generation sequencing in clinical oncology: next steps towards clinical validation. Cancers (Basel) 6(4):2296–2312. doi:10.3390/cancers6042296

Berkman PJ, Lai K, Lorenc MT et al (2012) Next-generation sequencing applications for wheat crop improvement. Am J Bot 99(2):365–371. doi:10.3732/ajb.1100309

Bianchi DW et al (2012) Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing. Obstet Gynecol 119(5):890–901
Bianchi DW et al (2014) DNA sequencing versus standard prenatal aneuploidy screening. N Engl J Med 370(9):799–808
Devers PL, Cronister A, Ormond KE et al (2013) Noninvasive prenatal testing/noninvasive prenatal diagnosis: the position of the National Society of Genetic Counselors. J Genet Couns 22(3):291–295. doi:10.1007/s10897-012-9564-0
Dondorp W, de Wert G, Bombard Y et al (2015) Non-invasive prenatal testing for aneuploidy and beyond: challenges of responsible innovation in prenatal screening. Summary and recommendations. Eur J Hum Genet doi:. doi:10.1038/ejhg.2015.56
Farrer RA, Kemen E, Jones JDG et al (2009) De novo assembly of the pseudomonas syringae pv. syringae B728a genome using illumina/solexa short sequence reads. FEMS Microbiol Lett 291:103–111
Hesselberth JR, Chen X, Zhang Z et al (2009) Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. Nat Methods 6:283–289
http://news.thermofisher.com/press-release/life-technologies/thermo-fisher-scientific-announces-listing-ion-pgm-dx-system-us-fda-
http://pediseq.research.chop.edu/
http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm375742.htm
https://www.genome.gov/10001772
https://www.genome.gov/sequencingcosts/
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945. doi:10.1038/nature03001
Korbel JO, Urban AE, Affourtit JP et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420–426
Krueger F, Kreck B, Franke A et al (2012) DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9(2):145–151. doi:10.1038/nmeth.1828
Krumm N, Sudmant PH, Ko A et al (2012) Copy number variation detection and genotyping from exome sequence data. Genome Res 22:1525–1532. doi:10.1101/gr.138115.112
Lebofsky R, Decraene C, Bernard V et al (2015) Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. Mol Oncol 9(4):783–790. doi:10.1016/j.molonc.2014.12.003
Ledford H (2008) The death of microarrays? Nature 455(7215):847. doi:10.1038/455847a
Lianos GD, Mangano A, Cho WC et al (2015) Circulating tumor DNA: new horizons for improving cancer treatment. Future Oncol 11(4):545–548. doi:10.2217/fon.14.250
Lo YMD, Corbetta N, Chamberlain PF et al (1997) Presence of fetal DNA in maternal plasma and serum. Lancet 350:485–487
Madoui MA, Engelen S, Cruaud C et al (2015) Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16(1):327
Mamanova L, Coffey AJ, Scott CE et al (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7(2):111–118. doi:10.1038/nmeth.1419
Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10(10):669–680
Rabbani B, Mahdieh N, Hosomichi K et al (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet 57:621–632
Ruike Y, Imanaka Y, Sato F et al (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. doi:10.1186/1471-2164-11-137

Sánchez-Quinto F, Lalueza-Fox C (2015) Almost 20 years of Neanderthal palaeogenetics: adaptation, admixture, diversity, demography and extinction. Philos Trans R Soc Lond B Biol Sci 370(1660):20130374. doi:10.1098/rstb.2013.0374

Sanger F, Nicklen S, Coulson AR (1977) Proc Natl Acad Sci U S A 74(12):5463–5467

Shyr D, Liu Q (2013) Next generation sequencing in cancer research and clinical application. Biol Proced Online 15(1):4. doi:10.1186/1480-9222-15-4

Song L, Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010(2):pdb.prot5384

Thomas T, Gilbert J, Meyer F (2012) Metagenomics – a guide from sampling to data analysis. Microb Inform Exp 2(1):3. doi:10.1186/2042-5783-2-3

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. doi:10.1038/nrg2484

Wang et al (2013) Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. Prenatal Diagnosis 33(7):662-6

Waterston RH, Lander ES, Sulston JE (2002) On the sequencing of the human genome. Proc Natl Acad Sci U S A 99(6):3712–3716

www.454.com

www.allseq.com

www.illumina.com

www.iontorrent.com

www.nanoporetech.com/

www.pacificbiosciences.com