Chapter 1

Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications

Jerzy K. Kulski

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61964

Abstract

Next-generation sequencing (NGS) technologies using DNA, RNA, or methylation sequencing have impacted enormously on the life sciences. NGS is the choice for large-scale genomic and transcriptomic sequencing because of the high-throughput production and outputs of sequencing data in the gigabase range per instrument run and the lower cost compared to the traditional Sanger first-generation sequencing method. The vast amounts of data generated by NGS have broadened our understanding of structural and functional genomics through the concepts of "omics" ranging from basic genomics to integrated systeomics, providing new insight into the workings and meaning of genetic conservation and diversity of living things. NGS today is more than ever about how different organisms use genetic information and molecular biology to survive and reproduce with and without mutations, disease, and diversity within their population networks and changing environments. In this chapter, the advances, applications, and challenges of NGS are reviewed starting with a history of first-generation sequencing followed by the major NGS platforms, the bioinformatics issues confronting NGS data storage and analysis, and the impacts made in the fields of genetics, biology, agriculture, and medicine in the brave, new world of "omics."

Keywords: Next-generation sequencing, tools, platforms, applications, omics

1. Introduction

Next-generation sequencing (NGS) refers to the deep, high-throughput, in-parallel DNA sequencing technologies developed a few decades after the Sanger DNA sequencing method first emerged in 1977 and then dominated for three decades [1, 2]. The NGS technologies are different from the Sanger method in that they provide massively parallel analysis, extremely high-throughput from multiple samples at much reduced cost [3]. Millions to billions of DNA

nucleotides can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that were used with Sanger sequencing [4]. The second-generation sequencing methods are characterized by the need to prepare amplified sequencing libraries before undertaking sequencing of the amplified DNA clones, whereas third-generation single molecular sequencing can be done without the need for creating the time-consuming and costly amplification libraries [5]. The parallelization of a high number of sequencing reactions by NGS was achieved by the miniaturization of sequencing reactions and, in some cases, the development of microfluidics and improved detection systems [6]. The time needed to generate the gigabase (Gb)-sized sequences by NGS was reduced from many years to only a few days or hours, with an accompanying massive price reduction. For example, as part of the Human Genome Project, the J. C. Venter genome [7] took almost 15 years to sequence at a cost of more than 1 million dollars using the Sanger method, whereas the J. D. Watson (1962 Nobel Prize winner) genome was sequenced by NGS using the 454 Genome Sequencer FLX with about the same 7.5x coverage within 2 months and for approximately 100th of the price [8]. The cost of sequencing the bacterial genome is now possible at about $1000 (https://www.nanoporetech.com), and the large-scale whole-genome sequencing (WGS) of 2,636 Icelanders [9] has brought some of the aims of the 1000 Genomes Project [10] to abrupt fruition.

Rapid progress in NGS technology and the simultaneous development of bioinformatics tools has allowed both small and large research groups to generate *de novo* draft genome sequences for any organism of interest. Apart from using NGS for WGS [11], these technologies can be used for whole transcriptome shotgun sequencing (WTSS) — also called RNA sequencing (RNA-seq) [12], whole-exome sequencing (WES) [13], targeted (TS) or candidate gene sequencing (CGS) [14–16], and methylation sequencing (MeS) [17]. RNA-seq can be used to identify all transcriptional activities (coding and noncoding) or a select subset of targeted RNA transcripts within a given sample [12], and it provides a more precise and sensitive measurement of gene expression levels than microarrays in the analysis of many samples [18–21]. In contrast to WGS, WES provides coverage for more than 95% of human exons to investigate the protein-coding regions (CDS) of the genome and identify coding variants or SNPs when WGS and WTSS are not practical or necessary [13]. Since the exome represents less than 2% of the human genome, it is the cost-effective alternative to WGS and RNA-seq in the study of human genetics and disease [13]. However, WGS may be preferred over WES because it provides more data with better uniformity of read coverage on disease-associated variants and reveals polymorphisms outside coding regions and genomic rearrangements [19, 22]. The analysis of the methylome by MeS complements WGS, WES, and CGS to determine the active methylation sites and the epigenetic markers that regulate gene expression, epistructural base variations, imprinting, development, differentiation, disease, and the epigenetic state [23–30]. The impact of NGS technology is indeed egalitarian in that it allows both small and large research groups the possibility to provide answers and solutions to many different problems and questions in the fields of genetics and biology, including those in medicine, agriculture, forensic science, virology, microbiology, and marine and plant biology.

http://dx.doi.org/10.5772/61964

The aim of this chapter is to provide an overview of the advances, applications, and challenges of NGS, starting with a history of first-generation sequencing followed by the major NGS platforms, the bioinformatics issues confronting NGS data storage and analysis, and the applications and challenges for biology and medicine in the world of "omic" expansion.

2. First-generation sequencing: A brief history

Twelve years after the publication of the Watson and Crick double-helix DNA structure in 1953 [31], the first natural polynucleotide sequence was reported [32]. It was the 77-nt yeast alanine tRNA with a proposed cloverleaf structure, although the anticodon, the three nucleotides that bind to the mRNA sequence, was not yet identified in the sequence [32]. It took 7 years to prepare up to 1 g of the tRNA from commercial baker's yeast by countercurrent distribution before fragmenting the RNA into short oligonucleotides with various RNase enzymes to reconstruct and identify the nucleotide residues using two-dimensional chromatography and spectrophotometric procedures [33]. At that time, scientists could sequence only a few base pairs per year, not nearly enough to sequence an entire gene. Nevertheless, despite the time-consuming and laborious nature of these very first sequencing methods that were developed for tRNA and other oligonucleotides, there was a flurry of RNA and DNA sequencing for the next 10 years that improved the sequencing procedures of fragmented DNA and provided new information on the sequences of more than 100 different tRNA. These initial labor-intensive sequencing efforts resulted also in the first complete genome sequence — the 3,569-nucleotide-long bacteriophage MS2 RNA, the lysozyme gene sequence of bacteriophage T4 DNA, and the 24-bp lac operator sequence [33–36]. This eventually led to the Maxam and Gilbert chemical degradation DNA sequencing method that chemically cleaved specific bases of terminally labeled DNA fragments and separated them by electrophoresis [37]. New data on how to sequence bacteriophage DNA by specific primer extension methods resulted in Sanger et al. [1] using primer-extension and chain-termination methods for sequencing polynucleotides longer than oligonucleotide lengths. Subsequently, the new Sanger DNA chain-termination sequencing method [1], known simply as the Sanger sequencing method, prevailed over the Maxam and Gilbert chemical degradation method [37] because of its greater simplicity and reliability and the use of fewer toxic chemicals and lower amounts of radioactivity. The first-generation automated DNA sequencers developed by Applied Biosystem Instruments (ABI) used the Sanger method with fluorescent dye-terminator reagents for single-reaction sequencing rather than the usual four separate reactions [34–36]. These sequencers were later improved by including computers to collect, store, and analyze the sequencing data [38]. The invention of the PCR technology [39] and thermal cyclers and the use of a heat-resistant enzyme such as Taq polymerase from *Thermus aquaticus* between 1985 and 1990 enabled the generation of random or specific sequences for *de novo* sequencing, filling gaps, and resequencing particular regions of interest [35]. The discovery of reverse transcriptase in 1970 [40, 41] led to the development of RNA sequencing using cDNA reverse transcribed from RNA. In 1991, Adams et al. [42] initiated a systematic cDNA

sequencing project using the Sanger method and the 373A DNA semiautomated sequencers to generate large batches of cDNA sequences with an average length of 397 bases, which they named "expressed sequence tags" (ESTs) and used as substrates and markers for RNA contig and transcriptome mapping. These improvements, together with the establishment of GenBank (http://www.ncbi.nlm.nih.gov/genbank) in 1982, resulted in the generation of hundreds of thousands of more DNA sequences throughout the 1980s, 1990s [34–36], and right up to the beginning of the new millennium, with the publication of the first draft sequence of the human genome [43, 44].

A sudden increase in the number of DNA and RNA sequences generated for GenBank between 1992 and 2004 (http://www.ncbi.nlm.nih.gov/genbank/statistics) resulted mostly from three main initiatives: the development of automated sequencers and the emergence of service providers, the industrialization and the establishment of sequencing centers and international consortiums, and the continued development of computing hardware and software to store and analyze nucleotide sequences. The automated-industrialized approach based on random or shotgun sequencing was initiated by The Institute for Genomic Research (TIGR) in Rockville, Maryland, and resulted in the publication of 337 new human genes and 48 homologous genes from other organisms [42]. By 1999, the TIGR venture generated 83 million nucleotides of cDNA sequence, 87,000 human cDNA sequences, and the complete genome sequences of two bacterial species, *Haemophilus influenzae* [45] and *Mycoplasma genitalium* [46]. This success was in part due to the development of the TIGR sequence assembler, an innovative computer program to assemble vast amounts of EST data [47]. By the end of 2001, the automated sequencers, such as the fully automated Prism 3700 with 96 capillaries that could produce 1.6×10^5 bases of sequence data per day, sequencing centers and international consortiums, such as the TIGR in the USA, the Sanger Centre in the United Kingdom, and RIKEN in Japan, produced the complete genomic sequences of the bacteria *E. coli* and *Bacillus subtilis*, the yeast *Saccharomyces cerevisiae*, the nematode *C. elegans*, the fruit fly *Drosophila melanogaster*, the plant *Arabidopsis thialiana*, and the human genome (see references cited by Stein [48]). Although sequencing was still hugely expensive and time consuming, Sanger sequencing was by then the dominant method. Pundits now placed DNA sequencing into a postgenomic era and predicted functional genomics, SNPs, and transcript arrays as the future of biological investigation [49, 50]. Indeed, after the establishment of the first Affymetrix and GeneChip microarrays in 1996, the decade saw a rapid growth in DNA array technology and applications for various gene expression studies in prokaryotes and eukaryotes [21, 51, 52]. Nevertheless, the outputs for genomic and/or RNA sequencing had neither finished nor slowed; new sequencing methods continued to emerge after 2005 to challenge the cost and supremacy of the Sanger dideoxy method [34–36]. These new methods became known as next-generation sequencing because they were designed to employ massively parallel strategies to produce large amounts of sequence from multiple samples at very high-throughput and at a high degree of sequence coverage to allow for the loss of accuracy of individual reads when compared to Sanger sequencing. These different approaches brought the cost of sequencing the genome down from $100 million in 2001 to less than $10,000 in 2014 [53].

3. Second-generation sequencing

A more detailed history of the development of the first- and next-generation sequencing platforms has been presented in a number of previous reviews [2–6, 11, 34–36, 54]. Table 1 outlines the basic features and performances of the common next-generation sequencing platforms. The basic characteristics of second-generation sequencing technology are the following. Shotgun sequencing of random fragmented genomic (fg) DNA or cDNA reverse transcribed from RNA is performed without the need for cloning via a foreign host cell: instead, linker and/or adapter sequences are ligated to the fgDNA or cDNA for construction of template libraries. Library amplification is performed on a solid surface or on beads while isolated within miniature emulsion droplets or arrays. Nucleotide incorporation is monitored directly by luminescence detection or by changes in electrical charge during the sequencing procedure. NGS generates many millions of nucleotide short reads in parallel in a much shorter time than by the Sanger sequencing method. The read types generated by NGS are digital and therefore enable direct quantitative comparisons. Either single or pair end reads can be obtained at fragment ends.

3.1. DNA and RNA library preparations for second-generation sequencing

The general workflow for second-generation sequencing is the preparation and amplification of libraries prepared from DNA or RNA samples, clonal formation, sequencing, and analysis [55–59]. Head et al. [55] have reviewed the methods and problems encountered for preparing NGS libraries for whole-genome sequencing, exome sequencing, target sequencing, RNA-seq, ChIP-seq, RIP-seq, and methylation sequencing (methyl-seq). Prior to library preparation, the genomic DNA is fragmented by acoustic shearing, sonication, or enzymatic digestion with DNase I or fragmentase and then labeled with adapters, tags, barcodes, and primers using established ligation and PCR methods. Alternatively, Illumina's fragmentation technology, called Nextera Tagmentation, can be implemented using a transposase enzyme to simultaneously fragment and insert adapter sequences into the ds DNA and thereby reduce sample handling and preparation time [57]. For targeted sequencing, the exomes or regions of interest within the fragmented DNA can be captured and enriched by probe-hybridization-capture kits or by PCR amplification with custom-designed primers. For RNA-seq of mRNA, polyA-RNA is isolated usually from total RNA or rRNA-depleted RNA and reverse transcribed to cDNA with reverse transcriptase and polyT or polyU primers before being treated much the same way as the fragmented genomic DNA. RNA sequencing libraries also can be created from immunoprecipitated RNA-binding proteins. To isolate noncoding RNAs (micro, small, and long) from total RNA, these sequences are selectively ligated to 3′ and 5′ adapters and reverse transcribed to cDNA. For methylation sequencing, the genomic DNA is reacted usually with bisulfite chemicals prior to library construction. On the other hand, ChIP-seq and RIP-seq use antibody capture to enrich the relevant sequences before preparation of the genomic DNA fragments for sequencing. In comparison to high-input gDNA libraries, the RNA and ChIP libraries may be limited by low cell numbers as starting material and consequently result in a low input of extracted DNA from the immunoprecipitated histones or DNA-binding proteins and in a limited sequence coverage.

Numerous DNA and RNA library kits and machines are available for the semiautomated or fully automated preparation of DNA libraries both for second- and/or third-generation sequencing. Some of these are GemCode from x10 (http://10xgenomics.com) and Raindrop's Thunderstorm (http://raindancetech.com) for all sequencing platforms, cBot for the Illumina platform [58], and Ion Chef and Ion OneTouch for the Ion Torrent platform [59]. All of these kits and auxiliary machines attempt to reduce workload and costs for the main platform sequencers. The DNA libraries are labeled with barcode sample tags, such as the multiplex identifier (MID) for Roche/454 sequencing, to enable the libraries to be pooled and therefore maximize the sequence output as a multiplex amplicon sequencing step for each sequencing run. After library construction, the DNA fragments are clonally amplified by emulsion PCR with microbeads [4, 6, 60] or by solid-phase PCR using primers attached to a solid surface [4, 61, 62] in order to generate sufficient single-stranded DNA molecules and detectable signal for producing sufficiently reliable sequencing data [54]. Roche 454, Life Technologies' SOLiD, and Ion Torrent platforms use emulsion PCR, whereas Illumina's HiSeq/MiSeq platforms use solid-phase PCR [4]. More recently, isothermal PCR amplification on a solid surface of a flow cell [62] was developed for the SOLiD 5500 W series of sequencing machines.

A problem with preparing sequencing libraries by PCR amplification is that PCR introduces GC bias, a major source of unwanted variation and errors in the sequencing coverage [63]. Using alternative methods to PCR amplification improves library complexity and the coverage of high GC regions and reduces the number of duplicate reads [64]. A number of different PCR-free library preparation kits are available commercially, such as NEXTflex PCR-Free from Bioo Scientific, Accel NGS 2S PCR-free library kit from Swift Biosciences, and the Illumina TruSeq DNA PCR-Free Sample Preparation Kit that uses ligation amplification for Illumina and other sequencing platform systems.

3.2. NGS platforms

The main features and performances of five commonly used second-generation sequencing technologies that have been reviewed in detail by others [2–4, 11, 36, 54] are shown in Table 1.

NGS platforms/company/max output per run	Read length per run (bp)	No. reads per run	Time (h or days)	Cost per 10^6 bases	Raw error rate (%)	Platform cost (USD approx.)	Chemistry
First generation							
Sanger/Life Technologies/84 kb	800	1	2 h	2400	0.3	95,000	Dideoxy terminator
Second generation							
454 GS FLX+/Roche/0.7 Gb	700	1×10^6	24/48 h	10	1	500,000	Pyrosequencing
GS Junior/Roche/70 Mb	500	1×10^5	18 h	9		100,000	Pyrosequencing
HiSeq/Illumina/1500 Gb	2x150	5×10^9	27/240 h	0.1	0.8	750,000	Reversible terminators
MiSeq/Illumina/15 Gb	2x300	3×10^8	27 h	0.13	0.8	125,000	Reversible terminators
SOLiD/Life Technologies/120 Gb	50	1×10^9	14 days	0.13	0.01	350,000	Ligation
Retrovolocity/BGI/3000 Gb	50	1×10^9	14 days	0.01	0.01	12×10^6	Nanoball/ligation

NGS platforms/company/max output per run	Read length per run (bp)	No. reads per run	Time (h or days)	Cost per 10^6 bases	Raw error rate (%)	Platform cost (USD approx.)	Chemistry
Ion Proton/Life Technologies/100 Gb	200	6×10^7	2–5 h	1	1.7	215,000	Proton detection
Ion PGM/Life Technologies/2 Gb	200	5×10^6	2–5 h	1	1.7	80,000	Proton detection
Third generation							
SMRT/Pac Bio/1 Gb	>10,000	1×10^6	1–2 h	2	12.9	750,000	Real-time SMS
Heliscope/Helicos/25 Gb	35	7×10^9	8 days	0.01	0.2	1.35×10^6	Real-time SMS
Nanopore/Oxford Nanopore Technologies/1 Gb	>5000	6×10^4	48/72 h	<1	34	1000	Real-time SMS
Electron microscopy/ZS	7200		14 h	<0.01		1×10^6	Real-time SMS
Genia nanopore (http://www.geniachip.com)							Real-time SMS

Table 1. Basic features and performances of NGS platforms. Sources are [4, 11, 20, 54, 115]. For comparison of the NGS outputs, the human genome has 3×10^9 bp or 3 Gb.

3.2.1. Roche 454 pyrosequencing

Roche 454 pyrosequencing by synthesis (SBS) was the first commercially successful second-generation sequencing system developed by 454 Life Sciences in 2005 and acquired by Roche in 2007 (http://www.my454.com). This technology uses sequencing chemistry, whereby visible light is detected and measured after it is produced by an ATP sulfurylase, luciferase, DNA polymerase enzymatic system in proportion to the amount of pyrophosphate that is released during repeated nucleotide incorporation into the newly synthesized DNA chain [2, 4, 6]. The system was miniaturized and massively parallelized using PicoTiterPlates to produce more than 200,000 reads at 100 to 150 bp per read with an output of 20 Mb per run in 2005 [6]. The upgraded 454 GS FLX Titanium system released by Roche in 2008 improved the average read length to 700 bp with an accuracy of 99.997% and an output of 0.7 Gb of data per run within 24 h. The GS Junior bench-top sequencer produced a read length of 700 bp with 70 Mb throughput and runtime of 10 to 18 h. The major drawbacks of this technology are the high cost of reagents and high error rates in homopolymer repeats. The estimated cost per million bases is $10 by Roche 454 compared to $0.07 by Illumina HiSeq 2000 [54]. A more serious challenge for those using this technology is the announcement by Roche that they will no longer supply or service the 454 sequencing machines or the pyrosequencing reagents and chemicals after 2016 [65].

3.2.2. Illumina (Solexa) HiSeq and MiSeq sequencing

Illumina (http://www.illumina.com) purchased the Solexa Genome Analyzer in 2006 and commercialized it in 2007 [66, 67]. Today, it is the most successful sequencing system with a claimed >70% dominance of the market, particularly with the HiSeq and MiSeq platforms. The Illumina sequencer is different from the Roche 454 sequencer in that it adopted the technology of sequencing by synthesis using removable fluorescently labeled chain-terminating nucleo-

tides that are able to produce a larger output at lower reagent cost [4, 6, 66]. The clonally enriched template DNA for sequencing is generated by PCR bridge amplification (also known as cluster generation) into miniaturized colonies called polonies [66]. The output of sequencing data per run is higher (600 Gb), the read lengths are shorter (approximately 100 bp), the cost is cheaper, and the run times are much longer (3-10 days) than most other systems [54]. Illumina provides six industrial-level sequencing machines (NextSeq 500, HiSeq series 2500, 3000, and 4000, and HiSeq X series five and ten) with mid to high output (120–1500 Gb) as well as a compact laboratory sequencer called the MiSeq, which, although small in size, has an output of 0.3 to 15 Gb and fast turnover rates suitable for targeted sequencing for clinical and small laboratory applications [68]. The MiSeq uses the same sequencing and polony technology such as the high-end machines, but it can provide sequencing results in 1 to 2 days at much reduced cost [54]. Illumina's new method of synthetic long reads using TruSeq technology apparently improves *de novo* assembly and resolves complex, highly repetitive transposable elements [69].

3.2.3. Sequencing by Oligonucleotide Ligation and Detection (SOLiD)

Supported Oligonucleotide Ligation and Detection (SOLiD) is a next-generation sequencer instrument marketed by Life Technologies (http://www.lifetechnologies.com) and first released in 2008 by Applied Biosystems Instruments (ABI). It is based on 2-nucleotide sequencing by ligation (SBL) [4, 6, 66]. This procedure involves sequential annealing of probes to the template and their subsequent ligation. Sequencers on the market today, such as the 5500 W series, are suitable for small- and large-scale projects involving whole genomes, exomes, and transcriptomes. Previously, sample preparation and amplification was similar to that of Roche 454 sequencing [66]. However, the upgrades to Wildfire chemistry have enabled greater throughput and simpler workflows by replacing beads with direct *in situ* amplification on FlowChips and paired-end sequencing [62]. The SOLiD 5500 W series sequencing reactions still use fluorescently labeled octamer probes in repeated cycles of annealing and ligation that are interrogated and eventually deciphered in a complex subtractive process using Exact Call Chemistry that has been well described by others [2, 36, 66]. The advantage of this method is accuracy with each base interrogated twice. The major disadvantages are the short read lengths (50–75 bp), the very long run times of 7 to 14 days, and the need for state-of-the-art computational infrastructure and expert computing personnel for analysis of the raw data.

3.2.4. DNA nanoball sequencing by BGI Retrovolocity

Complete Genomics (http://www.completegenomics.com) developed DNA nanoball sequencing (DNBS) as a hybrid of sequencing by hybridization and ligation [70]. Small fragments (440–500 bp) of genomic DNA or cDNA are amplified into DNA nanoballs by rolling-circle replication that requires the construction of complete circular templates before the generation of nanoballs. The DNA nanoballs are deposited onto an arrayed flow cell, with one nanoball per well sequenced at high density. Up to 10 bases of the template are read in 5′ and 3′ direction from each adapter. Since only short sequences, adjacent to adapters, are read, this sequencing format resembles a multiplexed form of mate-pair sequencing similar to using Exact Call Chemistry in SOLiD sequencing [2, 36, 66]. Ligated sequencing probes are removed, and a new

pool of probes is added, specific for different interrogated positions. The cycle of annealing, ligation, washing, and image recording is repeated for all 10 positions adjacent to one terminus of one adapter. This process is repeated for all seven remaining adapter termini. Although the developers have sequenced the whole human genome, the major disadvantage of DNBS is the short length of reads and the length of time for the sequencing projects. Claimed cost of the reagents for sequencing of the whole human genome is under $5000. The major advantage of this approach is the high density of arrays and therefore the high number of DNBs (~350 million) that can be sequenced. In 2015, the Chinese genomics service company BGI-Shenzhen acquired Complete Genomics and introduced the Retrovolocity system for large-scale, high-quality whole-genome and whole-exome sequencing with 50x coverage per genome and with the sample to assembled genome produced in less than 8 days [71]. Complete Genomics claims to have sequenced more than 20,000 whole human genomes over 5 years and published widely on the use of their NGS platform. They provide public access to a human repository of 69 genomes data and a cancer data set of two matched tumor and normal sample pairs at http://www.completegenomics.com/public-data/.

3.2.5. Ion torrent

Ion Torrent technology (http://www.iontorrent.com) was developed by the inventors of 454 sequencing [60], introducing two major changes. Firstly, the nucleotide sequences are detected electronically by changes in the pH of the surrounding solution proportional to the number of incorporated nucleotides rather than by the generation of light and detection using optical components. Secondly, the sequencing reaction is performed within a microchip that is amalgamated with flow cells and electronic sensors at the bottom of each cell. The incorporated nucleotide is converted to an electronic signal detected by the electronic sensors. The two sequencers in the market that use Ion Torrent technology are the high-throughput Proton sequencer with more than 165 million sensors and the Ion Personal Genome Machine (PGM), a bench-top sequencer with 11.1 million sensors. There are four sequencing chips to choose from [72]. The Ion PI Chip is used with the Proton sequencer, and the Ion 314, 316, or 318 Chips are used with the Ion PGM. The Ion 314 Chip provides the lowest reads at 0.5 million reads per chip, whereas the Ion 318 Chip provides the highest reads of up to 5.5 million reads per chip. The Proton sequencer provides a higher throughput (10–100 Gb vs. 20 Mb–1 Gb) and more reads per run (660 Mb vs. 11 Mb) than the PGM chips, but the read lengths (200–500 bp), run time (4–5 h), and accuracy (99%) are similar [54, 72]. Sample preparation for the generation of DNA libraries is similar to the one used for Roche 454 sequencing but can be simplified with the use of the Ion Chef system for automated template preparation and chip loading. The Ion Torrent chip is used with an ion-sensitive field-effect transistor sensor that has been engineered to detect individual protons produced during the sequencing reaction. The chip is placed within the flow cell and is sequentially flushed with individual unlabeled dNTPs in the presence of the DNA polymerase. Incorporation of nucleotide into the DNA chain releases H protons and changes the pH of the surrounding solution that is proportional to the number of incorporated nucleotides. The major disadvantages of the system are problems in reading homopolymer stretches and repeats. The major advantages seem to be the relatively longer

read lengths, flexible workflow, reduced turnaround time, and a cheaper price than those provided by the other platforms [54, 73].

4. Third-generation sequencing: Emerging technologies for single-molecule sequencing

Third-generation single-molecule sequencing technologies have emerged to reduce the price of sequencing and to simplify the preparatory procedures and sequencing methods [4, 74, 75].

4.1. Single-molecule real-time (SMRT) sequencing by pacific biosciences

Pacific Biosciences (http://www.pacificbiosciences.com) markets the PacBio RS II sequencer and the SMRT real-time sequencing system [74, 75]. SMRT sequencing is performed in SMRT cells that contain 150,000 ultra-microwells at a zeptoliter scale where one molecule of DNA polymerase is immobilized at the bottom of each well using the biotin-streptavidin system in nanostructures known as zero-mode waveguides (ZMWs). Once the template single-strand DNA is coupled with immobilized DNA polymerase, fluorescently labeled dNTP analogs are added and detected when the nucleotide is incorporated into the growing strand. CCD cameras continuously monitor the 150,000 ZMWs as a series of observed pulses that are converted into single molecular traces representing the template sequences. Since all four nucleotides are added simultaneously and measured in real time, the speed of sequencing is much increased compared to technologies where individual nucleotides are flushed sequentially. Although the reported accuracy was 99.3% initially with read lengths of about 900 bp [4], circularizing the template and sequencing it several times using a technology called SMRTbell templates provided longer reads and improved the accuracy to >99.999% [76, 77]. Once sequencing is initiated, the system's computational Blade Center performs real-time signal processing, base calling, and quality assessment. Primary analysis data, including trace and pulse data, read-length, distribution, polymerase speed, and quality measurement, is streamed directly to the secondary analysis software called SMRT Analysis that is capable of processing sequencing data in real time. The secondary analysis tools also include a full suite of tools to analyze single-molecule sequencing data for a broad range of applications.

4.2. Helicos sequencing by the genetic analysis system

The Helicos sequencing system was the first commercial implementation of single-molecule fluorescent sequencing [66, 78], marketed by the now bankrupt Helicos Biosciences. Today, the sequencing provider Seqll (http://seqll.com) sequences genomic DNA and RNA using the Helicos sequencing system and HeliScope single-molecule sequencers. DNA is sheared, tailed with polyA, and hybridized to a flow cell surface containing oligo-dT for sequencing-by-synthesis of billions of molecules in parallel. The polyA-tailed fragments of DNA molecules are hybridized directly to the oligo-dT50 bound on the surface of disposable glass flow cells. The addition of fluorescent nucleotides with a terminating nucleotide pauses the cyclical process until an image of one nucleotide for each DNA sequence has been captured, and then

the process is repeated until the fragments have been completely sequenced [75, 76]. This sequencing system is a combination of sequencing by hybridization and sequencing by synthesis using a DNA polymerase [79]. Sample preparation does not require ligation or PCR amplification and, therefore, largely avoids the GC content and size biases observed in other technologies [56]. The HeliScope sequencing read lengths range from 25 to over 60 bases, with 35 bases being the average. This method has successfully sequenced the human genome [80] to provide disease signatures in a clinical evaluation [81] and sequenced RNA to produce quantitative transcriptomes of tissues and cells [82].

4.3. Nanopore sequencing by Oxford Nanopore Technologies (MinION and PromethION)

Oxford Nanopore Technologies provides the latest single-molecule sequencing system [83, 84]. The MinION Mkl is a portable handheld device for DNA and RNA sequencing that attaches directly to a laptop/computer using a USB port, whereas the PromethION is a small bench-top system. Nanopore sequencing uses pores formed from proteins, such as haemolysin, a biological protein channel system in *Staphylococcus aureus* [85]. The idea behind DNA and RNA sequencing using nanopores is that the conductivity of ion currents in the pore changes when the strand of nucleic acid passes through it [83]. The flow of ion current depends on the shape of the molecule translocating through the pore. Since nucleotides have different shapes, each nucleotide is recognized by its effect on the change of the ionic current [86]. The key advantage of this approach is that sample preparation is minimal compared to second-generation sequencing methods, and long read lengths can be obtained in the kbp range. In addition, there are no amplification or ligation steps required before sequencing. The main problem with this technology is the requirement to optimize the speed of DNA translocation through the nanopore to ensure reliable measurement of the ionic current changes and reduce the high error rates of base calling [83–86]. At this time, Oxford Nanopore Technologies is in the beta testing phase, and users are required to join the MinION Access Programme and pay a fee of $1000 [83] to access a MinION starter pack (3 MinION MkI flow cells, a Nanopore sequencing kit, and a wash kit). Laver et al. [87] have assessed the performance of an earlier version of the MinION sequencing device and concluded that "the MinION is an exciting prospect; however, the current error rate limits its ability to compete with existing sequencing technologies, though we do show that MinION sequence reads can enhance contiguity of *de novo* assembly when used in conjunction with Illumina MiSeq data." They resequenced three bacterial genomes and estimated the error rate to be 38.2%, with mean and median reads of 2 and 1 kb, respectively, and with the longest single read of 98 kb. The low depth of coverage provided by the present nanopore technology is a possible barrier to accurate eukaryotic genome sequencing at the moment. Nevertheless, these are not intangibles and nanopore nucleic acid sequencing is envisaged to include methylation and direct RNA sequencing in the near future [83].

4.4. NGS by electron microscopy

The sequence of long, intact DNA molecules can be visualized and identified by using electron microscopy. The first report on the successful application of electron microscopy for NGS was

for the partial sequencing of DNA base pairs within intact DNA molecules using synthesized genomes of 3.3 and 7.2 kb length that were sequenced by enzymatically incorporating modified bases that contained atoms of increased atomic number and allowed for the direct visualization and identification of individually labeled bases [88]. In this sequencing process, the double strands of the DNA sample are separated into single strands using common enzymes and reactions. Then, the single-stranded DNA is labeled by PCR using dNTPs attached to heavy-atom metal labels that can be separated into identifiable electron microscope-generated images showing large black dots, small black dots, and large gray dots along the DNA molecule linearized by ZSG threading. Standard image-based technologies perform the reads and analysis of the labeled DNA using image analysis software that provides sequence data in real time. The sequenced molecules are reads in the range of 5 to 50 kb in length that are useful for *de novo* genome assembly and for analysis of full haplotypes and copy number variants. The company ZS Genetics (http://www.zsgenetics.com) offers a service to provide accurate, long-read, single-molecule DNA sequences using the NGS electron microscopy platform.

5. NGS service providers

Researchers who cannot afford to purchase NGS machines at prices varying between $80,000 and over 1 million USD (depending on the platform) plus the many add-ons, computing requirements, and infrastructural changes, instead, might consider using one of the many available sequencing service providers. For example, Novogene, which was founded in Beijing in 2011 and now is located also in Great Britain and the USA, provides NGS for human, animal, plant, and microbe applications using Illumina MiSeq, HiSeq, and X platforms for whole-genome *de novo* sequencing and resequencing, exome sequencing, targeted sequencing, transcriptomics for mRNA and small RNA, and metagenomics. Similarly, the South Korean company Macrogen provides all the NGS services using Illumina platforms as well as epigenome sequencing for methylations by bisulfite conversion, methyl-CpG binding domain, or chromatin immunoprecipitation. Prices may vary between $500 and $2,000 USD per sample depending on the sequencing project and the project workflow from sample preparation to bioinformatics analysis (https://www.scienceexchange.com). Table 2 lists some of the service providers, and others can be accessed at http://omicsmaps.com.

Service provider	Platforms	DNA sequencing (TS WG WES)	RNA-seq	Methyl-seq	Web address
BGI	All	+++	+	+	bgiamericas.com
Novogene	Illumina	+++	+	+	novogene.com
Macrogen	Illumina	+++	+	+	macrogen.com
	Ion Torrent	+++	+	+	
CD Genomics	Illumina	+++	+	+	cd-genomics.com
	Ion Torrent	+++	+	+	

Service provider	Platforms	DNA sequencing (TS WG WES)	RNA-seq	Methyl-seq	Web address
	PacBioRS II	+ + +	+	+	
	CEA**			+	
SeqWright Genomic	Illumina	+ + +	+	+	seqwright.com/researchservices
	Ion Torrent	+ + +	+	+	
	Roche 454	+ + +	+		
EpigenDx	Ion Torrent			+	epigendx.com
Centrillion Genomic	Illumina	+ + +	+		centrillionbio.com
NXT-DX	Illumina		+	+	nxt-dx.com
AGRF***	Illumina	+ + +	+	+	agrf.org.au
	CEA**			+	
Broad Institute	Illumina	+ + +	+	+	genomics.broadinstitute.org
Illumina	Illumina	+ + +	+	+	illumina.com
Exiqon	Illumina		+		exiqon.com
SEQLL	Helicos	+ + +	+	+	seqll.com
Eurofins Genomics	Illumina	+ + +	+		eurofinsgenomics.eu
	Roche 454	+ + +			
	Ion Torrent	+ - +			
	PacBioRS II	+ - +			
Millennium Science	PacBioRS II	+ + +		+	mscience.com.au/view/
Oxford Nanopore Technologies	MINion	+ + +	+		nanoporetech.com
Complete Genomics	Nanoball arrays	+ +			completegenomics.com

Table 2. NGS service providers. In the DNA sequencing column, TS is targeted sequencing, WG is whole-genome sequencing, and WES is whole-exome sequencing. *RNA-seq includes whole transcriptome, mRNA, long, small, and microRNA sequencing. **Methyl-seq (methylation sequencing) or epigenetic analysis is usually performed by bisulfite sequencing and either NGS or capillary electrophoresis analysis (CEA). Other analyses such as MBD, MeDIP-seq, or ChIP-seq may be provided. Helicos and PacBio platforms also enable the detection of methylation sites. ***AGRF = Australian Genomic Research Facility. Most of the listed service providers also may perform sample and library preparation, Sanger sequencing, specialist genotyping, data analysis, and bioinformatics service. Other service providers can be accessed via the High-Throughput Sequencing Map site at http://omicsmaps.com.

6. Performance of NGS platforms and sequencing errors

All NGS systems produce unique sequencing errors and biases that need to be identified and corrected. The major sequencing errors are largely related to high-frequency indel polymorphisms, homopolymeric regions, GC- and AT-rich regions, replicate bias, and substitution errors [89–91]. While the PGM quality scores underestimate the base accuracy, the Roche 454

quality scores tend to overestimate the base accuracy. A key consideration for generating high-quality, unbiased, and interpretable data from next-generation sequencing studies is to achieve sufficient sequence depth and coverage for statistical certainty. Low sequencing depth can contribute to high error rates stemming from base calling and mapping errors, which in turn can affect the statistical significance for identifying true genotypes, nucleotide variants, and single nucleotide polymorphism. Increased depth of coverage can help sequence alignment mapping to differentiate between true variants and errors, although it might not resolve errors due to assembly gaps. Good sequence library preparation is paramount to producing good sequence depth and coverage. A number of different library methods are available to achieve this goal depending on the NGS applications [55]. Sims et al. [92] reviewed in critical detail the guidelines and precedents for optimal sequencing depth and coverage in regard to sequencing genomes, exomes, transcriptomes, methylomes, and epigenomes by chromatin immunoprecipitation and sequencing and/or chromosome conformation capture.

No single study has compared the performance of all the available NGS platforms simultaneously using the same control genomic sequences. However, a comparison of three bench-top sequencers, the Roche GS Junior, the Illumina MiSeq, and Ion PGM, revealed large differences in cost, sequence capacity, and performance outcomes of genome depth, stability of coverage and read lengths, and quality for sequencing bacterial genomes [54, 93]. Most sequencing errors arose with indel polymorphisms, GC-rich regions, and homopolymeric regions. Overall, the two laboratories concluded that all the machines had certain limitations that needed to be taken into account when designing sequencing experiments [54, 93]. In a comparison of bacterial genome sequencing between PacBio, Ion Torrent, and three Illumina machines (MiSeq, GAIIx, and HiSeq 2000), the sequencers all provided high accuracy for GC-rich, neutral, and moderately AT-rich genomes [94]. The main exception was the poor coverage in the extremely AT-rich region of *Plasmodium falciparum* with a single 316 chip for the Ion Torrent PGM that resulted in no coverage for 30% of the genome. In this study, PacBio generated the longest reads but produced the least accurate SNP detection and the highest error rate of 13% compared to 1.78% for Ion Torrent and less than 0.04% for the Illumina platforms. In a different comparison, the performance of whole-genome sequencing platforms Illumina's HiSeq2000, Life Technologies' SOLiD 4 and 5500xl SOLiD, and Complete Genomics' sequencing system were evaluated for their ability to call SNVs and to evenly cover the genome and specific genomic regions [95]. The authors concluded that all the platforms had their shortfalls with a pronounced GC bias in GC-rich regions and false-positive rates and that the best solution is to integrate the sequencing data from the four different platforms because it combined the strengths of different technologies. In an analysis of bacterial CREBBP exons, three different NGS platforms appear to have worked comparably well for targeted exomic sequencing with the percentage of total read numbers aligned to a reference sequence resulting in 99.8% for Roche 454, 98.1% for Illumina MiSeq, and 90.7% for Ion Torrent PGM sequence reads [96]. However, the Illumina MiSeq data showed the highest substitution error rate, whereas the PGM data revealed the highest indel error rate. On the other hand, there was little difference between the Junior Roche and the Ion PGM platforms for "in phase" sequence genotyping of HLA loci, and either platform could be used with excellent results [16]. In this case, the lower cost of reagents and a slightly quicker turnaround time favored the Ion PGM platform [97].

Five sequencing platforms, Illumina HiSeq, Ion PGM, Ion Proton, PacBio RS, and Roche 454, were tested in a comparative evaluation of RNA-seq reproducibility using reference RNA standards at 19 laboratory sites [20]. The results showed high intraplatform and interplatform concordance for expression measures across the deep-count regions but highly variable reproducibility for splice junction and variant detection between all platforms. Despite fewer bases sequenced, the Proton, PGM, and 454 platforms detected more known junctions compared to Illumina HiSeq.

7. Bioinformatics: DNA and RNA data analysis and storage

Bioinformatics is a major rate-limiting step for NGS technology with respect to overcoming the growing challenges of storage, analysis, and interpretation of NGS data [98–100]. There are at least four tiers of nucleotide sequence analysis to consider when using the NGS platforms [98–104]. The first is generation of sequence reads using the software integrated within the sequencing instruments that convert the raw signals into base calling with short reads of nucleotide sequences and associated quality scores. The second is the alignment and assembly of contigs and scaffolds and variant detection. The third is annotation, data integration, and visualization of the assembled sequence. The fourth is the amalgamation of all the data from the different NGS platforms into a single, coherent, bioinformatic output with accessible links and tools for general and particular biological or forensic interest. The Internet-web addresses to source the bioinformatics tools and databases for NGS data analysis from the original raw sequencing data to functional biology can be obtained from the following references [99–104] and Table 3.

The raw sequencing signals produced by the manufacturer's sequencing machine or system are converted into nucleotide bases of short read data (base calling) with base quality scoring using the system's FASTQ format or the native raw data file formats (Illumina, SFF, HDF5, CG, or SOLID). Storage of raw signal (image) and sequencing data as short read archives in the FASTQ format or native raw data file formats is a problem in regard to computing resources for many research sequencing laboratories and commercial service providers. Thus, the conversion of FASTQ files to the more compact Sequence Alignment Map (SAM) format and its compressed Binary Alignment Map (BAM) format is recommended because it is easier to read and process for later bioinformatics analysis [99, 102]. The safe storage of the original raw sequences is important for bioinformatics analysis and corrections because it is the source of the initial sequencing errors that are either filtered out or left within the final assembled sequence. Quality checks are necessary to remove reads with low phred levels, sequence errors, and sequences such as primers, vectors, adapters, tags, and tails that were introduced experimentally during the preparation of the sequencing libraries [101]. Errors or biases associated with raw reads from the Illumina, Roche, and SOLiD platforms are mainly fluorophore-dependent errors, whereas the non-fluorophore platforms such as Ion Torrent produce their own unique errors and biases [99, 101]. Therefore, many different signal and image detection programs and base calling algorithms still need to be developed and tested in an attempt to improve the accuracy of base calling [101}. The raw sequence data (a mixture of raw files and

other metadata) from the NGS technologies can be submitted to the NCBI Sequence Read Archive database for DNA studies and to Gene Expression Omnibus and ArrayExpress for mRNA-seq or ChIP-Seq studies in order to receive a database accession number and to reference the raw sequence data in scientific publications [105]. The Sequence Read Archive (SRA) at NCBI also provides a fee-free, downloadable SRA computing toolkit to read the raw graphs and files from the different NGS platforms and to convert between different file formats (Table 3). Archive files in the SRA format (.sra) are converted into the FASTQ or SAM/BAM formats for input to downstream analysis using software programs (Table 3) to undertake the second tier analysis of sequence alignment (spliced and genomic), assembly, and variant detection.

The requirement for sequence alignment and variant detection at the second tier of bioinformatics depends on the complexity of the NGS project. Small sequence reads from small genomes (e.g., viruses) are less complex and easier to compute and align and assemble than the many more reads generated from large genomes of mammals or higher plants. The transfer of the preedited DNA data in the correct format to alignment and variant detection software is generally straightforward and there are many free and commercial software packages available to perform these tasks [99–104]. As often is the case, a single package does not suit all analytical requirements. There may need to be a degree of interchange and testing to find the best solutions as well as using appropriate and informative controls for standardization and normalization. Schlotterer et al. [104] have reviewed programs for genotype and SNP calling. ANGSD is a new multithreaded program suite that was developed recently to perform association mapping, population genetic analyses (population structure measures, allele frequency for cases and controls, admixture, and neutrality tests), SNP discovery, and genotype calling using the raw sequence data and genotype likelihoods in NGS data of human DNA samples for the 1000 Genomes Project [106].

The alignment of sequences to provide long assemblies (contigs and/or scaffolds) may take two different paths. One is comparative mapping of short reads aligned to reference sequences and the other is *de novo* assembly of overlapping reads [101]. The accuracy of *de novo* assembly can be confirmed or improved by integrating it with comparative alignment mapping to reference genomic sequences. Sequencing assemblers may employ different graph construction algorithms and preprocessing and postprocessing filter computations to flag, correct, or eliminate sequencing errors with no single computation solution. Some genome assemblers forgo the preprocess filtering step and they all differ in their ease of use, in the accuracy, efficiency, and quality of assembly, ability to fill gaps, and differentiate between error driven variants and true variants or SNPs and in the detection and elimination of repeats and sequencing errors [99]. According to El-Metwally et al. [99], an ideal assembler should have a set of layers with clearly defined inputs, communication output messages to facilitate the development of innovative, interactive, independent assemblers using the SAM/BAM file formats and the language of FASTG (http://fastg.sourceforge.net) for the next-generation environment. Another way to improve the quality of sequencing and assembly is to adopt a hybrid approach by using two or more different sequencing platforms and assembly software. A new software package *anytag* that fills gaps between paired-end reads to generate near-error-

free contigs of up to 190 kb appears to be a fivefold improvement over existing *de novo* genome assemblers such as *soap* and *Newbler* [107].

In a recent evaluation of the most commonly used *de novo* genome assemblers to assemble the genomes of three vertebrate species (snake, bird, and fish) by Assemblathon, the authors recommended not to trust the results of any single assembly, nor place too much faith in a single metric of quality or accuracy, but instead to choose an assembler that excels in the area of interest and expectation to provide sufficient coverage, continuity, and error-free bases [108]. End users were reminded that the use of the assembly tools is not straightforward and that they should first gain considerable familiarity with the computing hardware and software and become aware of the "ease of installation, use, and management" of each assembly tool. Many problems with *de novo* genome assembly remain inherent with recognizing and evaluating highly heterozygous and repetitive regions, segmental duplications, and sequencing errors and gaps. This is complicated further by the different read lengths, read counts, and error profiles that are produced by different NGS technologies. In addition, most assembled genomic sequences in publicly accessible databases are at the level of or below a standard draft (minimum standards for submission to public databases) rather than a "high-quality draft" assembly that is completed to at least 90% of the expected genome size.

The third tier of bioinformatics is to annotate, transcribe, and translate the genomic sequences to a higher informatics level, such as defining gene exon coding (CDS) and noncoding (5′ noncoding, introns, and 3′ terminal end) untranslated regions (UTRs), alternate transcript isoforms, signal peptides, repeat elements, and other nontranscribed regions such as viral integration sites and chromosomal common fragile sites [103]. Genomic sequences of prokaryotes are a thousand times smaller and less complex than those of eukaryotes and consequently are easier to assemble and annotate. A typical methodology for prokaryote annotation suggested by the National Pathogen Data Resource to annotate 1000 genomes is to first submit the genomic sequence to the Rapid Annotation Server (RAST) at the Argonne National Laboratory and receive back the protein-encoded genes (CDS), the RNA-encoded genes (tRNAs and rRNAs), and identified subsystems such as metabolic pathways, complex structures, and phenotypes (Table 3). This initial annotation should then be reanalyzed in detail to find discrepancies between the sequence and the translation using any other public or commercial genomic tools to fix miscalled genes and variants, frameshifts, insertion sequences, and pseudogenes. The public web server CRISPRfinder detects and annotates the bacterial CRISPRs and tandem repeat sequences that may impact on genes and pseudogenes (Table 3). After the reanalysis and final fixes, the annotated and curated genome should be rerun through RAST to update the subsystems output. Other useful web-based microbial annotation servers can be accessed at MicroScope, BASys, and NCBI's Prokaryotic Genome Annotation Pipeline (PGAP), with additional software provided at Prokka (Table 3). A typical prokaryotic genome annotation process is outlined at NCBI (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/).

Eukaryote genome annotation is more complex and challenging than prokaryote genome annotation. In an overview of the available tools and best practices for eukaryotic genome annotation, Yandell and Ence [103] pointed to five basic categories of annotation software: (1)

ab initio and evidence-drivable gene predictors; (2) EST, protein, and RNA-seq aligners and assemblers; (3) choosers and combiners; (4) genome annotation pipelines; and (5) genome browsers for curation. A typical eukaryotic genome annotation pipeline is outlined by NCBI at http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/. The essential first step for eukaryote genome annotation and gene determination is to identify and mask repeat elements (microsatellites, retrotransposons, and transposons) using RepeatMasker, Censor, or WindowMasker (Table 3). Without the initial masking step, the repeats would seed millions of spurious BLAST alignments and create incorrect gene annotations and corrupt the genome annotation with artifacts and false metadata. After masking, the annotation pipeline includes the following steps: transcript, RNA-seq read, protein/domain alignments; guided/*ab initio* gene model predictions; curated genomic sequence alignments; selection of the best evidence based models; gene naming and locus typing; assignment of GeneIDs; and annotation of small RNAs. In addition, there are the special considerations such as annotation of multiple assemblies and updated assemblies before the annotated products can obtain an Annotation Release number and a release date for availability in various NCBI resources, including the databases for nucleotides, proteins, BLAST, gene, Map Viewer, and FTP sites. Other websites and tools considered important for eukaryote annotation are BUSCO for assessing the "core" eukaryote genes, Babelomics for the functional analysis of transcriptomic and genomic data, the PASA and MAKER tools for updating annotations with RNA-seq data, and other data and information (Table 3). The annotated and mapped data can then be integrated, visualized, and presented at a fourth tier of bioinformatics with genome browsers such as those displayed at UCSC, Ensembl, JBrowse, Web Apollo (Table 3), and others such as Genome Maps [109]. The new Emsembl 2015 provides an up-to-date genomic interpretation system for annotations, query tools, and access methods for chordates and key model organisms [110].

Gene ontology is a bioinformatics initiative that provides (a) defined terms representing gene product properties and pathways covering biological domains such as cellular components, molecular function, and biological processes with their various subcategories and (b) functional annotation tools to find functions for large gene lists. It sits somewhere between the third tier (annotation) and the fourth tier of bioinformatic analyses and structures. The first major Gene Ontology (GO) project was founded in 1998 to address a need for standard filtered descriptions of gene products across different databases. GO is a collaborative effort that started between three model organism databases, FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD) but now incorporates many databases for plant, animal, and microbial genomes. The GO Contributors page lists all member organizations (http://geneontology.org/page/go-consortium-contributors-list). Some other ontology providers among many are the Open Biological and Biomedical Ontologies (OBBO), Reactome, DAVID, and the KEGG Pathway database (Table 3).

NGS manufacturers provide their own unique software for the first tier analysis to process the raw acquisition data and produce read files that contain high-quality consensus reads for the draft assemblies. However, only a few have attempted to include all three tiers of nucleotide sequence analysis into a fourth tier that is an easily accessible single integrated package. Illumina has provided the BaseSpace genomics cloud-computing program for integrated data

http://dx.doi.org/10.5772/61964

storage and analysis (Table 3). This cloud storage and analysis program permits instrument integration with sequence analysis viewing and access to a wide range of software applications to align, assemble, and analyze reads and variants for RNA and DNA. These apply to various workflows, including basic analyses for prokaryotic and eukaryotic genomics and transcriptomics, metagenomics, and for more specialist interests such as detection and analysis of tumor variants, haplotype analysis, pathways and networks, forensic profiles, and many others, too numerous to list here. In comparison, Ion Torrent has storage devices and servers with a web browser driving the Torrent Suite Software (Table 3) on computers attached to their respective sequencing instruments. The manufacturer's software can be used to preprocess the DNA sequencing read data before transferring the preedited data onto other analytical software systems that are either provided by the manufacturer (vendor) or obtained from elsewhere. The National Center for Biotechnology Information (NCBI) is an example of a fourth tier bioinformatics provider (Table 3) that is a free, one-stop shop for DNA and RNA sequence data, analysis, and information. There are direct links at NCBI to 65 accessible databases, 35 download sites (for databases, tools, and utilities), 17 public submission portals, and 60 computing tools for sequence and data analysis, reports, and tutorials. In addition, NCBI is a resource for books and journals through its online library and the PubMed webpage.

Program	**Website**
1. Aligner, assembly, and postassembly tools	
MUMmer aligner	http://mummer.sourceforge.net
Bowtie aligner	http://bowtie-bio.sourceforge.net/index.shtml
TopHat RNA-seq aligner	https://ccb.jhu.edu/software/tophat/index.shtml
Anytag aligner	http://sourceforge.net/projects/anytag/files/anytag2.0/
Soap *de novo* assembler	http://soap.genomics.org.cn/soapdenovo.html
Allpaths-LG assembler	http://www.broadinstitute.org/software/allpaths-lg/blog/
Celera assembler	http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page
Velvet assembler	https://www.ebi.ac.uk/~zerbino/velvet/
SPAdes assembler	http://bioinf.spbau.ru/spades
Galaxy tools	https://usegalaxy.org
Genomic tools	http://molbiol-tools.ca/Genomics.htm
BaseSpace Illumina	https://basespace.illumina.com/home/sequence
Torrent Suite Software	http://www.lifetechnologies.com/torrentsuite
RATT: rapid annotation transfer tool	http://ratt.sourceforge.net
2. Prokaryote annotation web servers	
RAST	http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/RapidAnnotationServer
	http://rast.nmpdr.org

Program	Website
CRISPRfinder	http://crispr.u-psud.fr/Server/CRISPRfinder.php
Mreps	http://bioinfo.lifl.fr/mreps/mreps.php
MicroScope	https://www.genoscope.cns.fr/agc/microscope/home/index.php
BaSys	https://www.basys.ca
PGAP	http://www.ncbi.nlm.nih.gov/genome/annotation_prok/
Prokka	http://www.vicbioinformatics.com/software.prokka.shtml
3. Eukaryote annotation web servers	
NCBI pipeline	http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
RepeatMasker	http://www.repeatmasker.org/
Censor	http://www.girinst.org/censor/
WindowMasker	http://nebc.nerc.ac.uk/bioinformatics/docs/windowmasker.html
CEGMA tool	http://korflab.ucdavis.edu/datasets/cegma/
BUSCO	http://busco.ezlab.org
PASA	http://pasapipeline.github.io
MAKER	http://www.yandell-lab.org/software/maker.html
Babelomics	http://www.babelomics.org
4. Archives and databases	
DDBJ	http://www.ddbj.nig.ac.jp
EMBL	http://www.embl.org
GenBank	http://www.ncbi.nlm.nih.gov/genbank/
REPBASE	http://www.girinst.org
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi
dbGAP	http://www.ncbi.nlm.nih.gov/gap
Complete Genomics data	http://www.completegenomics.com/public-data/
SRA	http://www.ncbi.nlm.nih.gov/sra
OMIM	http://www.ncbi.nlm.nih.gov/omim
COSMIC	http://cancer.sanger.ac.uk/cosmic
ENCODE	https://www.encodeproject.org
GTEx	http://www.gtexportal.org
FANTOM	http://fantom.gsc.riken.jp
Roadmap epigenomics	http://www.roadmapepigenomics.org
Blueprint epigenomics	http://www.blueprint-epigenome.eu
Regulome DB	http://regulomedb.org

Program	Website
ExPASy proteomics	http://www.expasy.org/proteomics/protein-protein_interaction
PRIDE proteomics	http://www.ebi.ac.uk/pride/archive/
FAME metabolomics	http://f-a-m-e.fame-vu.cloudlet.sara.nl
Metabolomexpress	https://www.metabolome-express.org
MetaboAnalyst	http://www.metaboanalyst.ca
AromaDeg	http://aromadeg.siona.helmholtz-hzi.de
EGA phenome	https://www.ebi.ac.uk/ega/home
GOLD	https://gold.jgi-psf.org
MG-RAST	https://metagenomics.anl.gov
ViralZone	http://viralzone.expasy.org
UCNEbase UC elements	http://ccg.vital-it.ch/UCNEbase/
UCbase 2.0 UC elements	http://ucbase.unimore.it/
DEG database	http://www.essentialgene.org
PhylomeDB	http://phylomedb.org/
Compara GeneTrees	http://asia.ensembl.org
TreeFam	http://treefam.genomics.org.cn
PANTHER	http://pantherdb.org
FATCAT	http://phylogenomics.berkeley.edu
HOGENOM database	http://doua.prabi.fr
5. *Gene ontology databases and tools*	
Gene Ontology Consortium	http://geneontology.org
OBBO	http://www.obofoundry.org, http://obofoundry.github.io
Reactome	http://www.reactome.org/
DAVID 6.7	https://david.ncifcrf.gov/
KEGG Pathway database	http://www.genome.jp/kegg/pathway.html
6. Genome browsers, projects, and fourth tier providers	
Kbase	http://kbase.us/glossary/systems-biology/
Earth Microbiome Project	http://www.earthmicrobiome.org
Terragenome Project	http://www.terragenome.org
Tara Oceans Project	http://ocean-microbiome.embl.de/companion.html
MetaHit project	http://www.metahit.eu
Vertebrate Genome 10K	http://genome10k.org
Human Microbiome	http://hmpdacc.org

Program	Website
Personal Genome Project	http://www.personalgenomes.org
1000 Genomes Project	http://www.1000genomes.org
HapMap	http://hapmap.ncbi.nlm.nih.gov/
UCSC browser	https://genome.ucsc.edu
Ensembl browser	http://www.ensembl.org
Jbrowse browser	http://jbrowse.org
Web Apollo browser	http://genomearchitect.org
NCBI mapview	http://www.ncbi.nlm.nih.gov/projects/mapview/
NCBI resources	http://www.ncbi.nlm.nih.gov/guide/all/ - tab-all_
KEGG	http://www.genome.jp/kegg/
7. Optical mappers	
BioNano mapper	http://www.bionanogenomics.com
Whole-Genome Mapping	http://opgen.com/genomic-services/what-is-whole-genome-mapping
8. NGS and bioinformatics software providers and biological databases	
Omicsmap for NGS	http://omicsmaps.com/
The NGS WikiBook	http://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)
The Sequencing Marketplace	http://allseq.com
Genomeweb	https://www.genomeweb.com
Bioinformatic software	http://seqanswers.com/wiki/Software/list
	https://en.wikipedia.org/wiki/List_of_open-source_bioinformatics_software
	http://bioinformaticssoftwareandtools.co.in
Bioinformatics Web	http://www.bioinformaticsweb.net
Biological databases	https://en.wikipedia.org/wiki/List_of_biological_databases
Applied Bioinformatics	http://www.appliedbioinformatics.com.au

Table 3. Useful websites for NGS tools, browsers, portals, providers, and online databases.

8. Impact and applications of NGS: Opening the doors into the world of "omics"

All hereditary information is contained within the structure, organization, and function of an organism's genome. The continual emergence of many new public bioinformatics databases (Table 3) on the World Wide Web demonstrates and reflects the impact of NGS on the life sciences and our need to constantly develop new methods to interrogate and decode hereditary

http://dx.doi.org/10.5772/61964

information in and around DNA (or RNA for some viruses) and its nucleotide sequences (http://www.bioinformaticsweb.net).

Although genomics is a relatively young field, arguably starting sometime between 1976 with the publication of the bacteriophage MS2 RNA genome [111] and 1986 when the word "genomics" was first used [112], it already has made an enormous impact on the life sciences. The term "genomics" coined by Thomas Roderick in 1986 encompassed the structure and function of genes, and comparative genomics elucidated the hereditary relationships and evolution within and between different species [112]. Since the advent of NGS, the meaning of "genomics" has been narrowed more towards mapping the structure and organization of genomes and differentiating between *de novo* sequences, resequenced genomes, exonic or targeted sequences, and metagenomic sequences. The other implied meanings of "genomics" are attributed now to the suffix "-omics," added to integrated fields undertaken on a large or genome-wide scale such as transcriptomics, haplomics, methylomics, epigenomics, proteomics, metabolomics, nutrigenomics, physiomics, evolomics, epidemiomics, systeomics, personomics, multinomics, etc. [113]. Thus, NGS broadens our understanding of structural and functional genomics through the concepts of "omics" to provide new insight into the workings and meaning of genetic conservation and diversity of living things (http://www.nature.com/omics/index.html). It is more than ever about how different organisms use genetics and molecular biology to survive and reproduce with and without mutations, disease, and diversity within their own life cycles and within their population networks and changing environmental conditions.

8.1. Genomics

A detailed organizational analysis and an understanding of the full landscape of a genome are possible only after *de novo* whole-genome shotgun sequencing and annotation has been performed [11]. WGS has had an enormous impact on viral, bacterial, and archaeal genomics [114–117]. Some of these successes are provided in the metagenomics section (see section *8.5*). Others have reviewed the impact of WGS and genomics on fungi [118, 119], algae [120], animals [121, 122], and humans [10, 13, 123–127]. WGS has become increasingly easier, faster, and cheaper because of technological improvements and the availability of hundreds of sequenced genomes that can be used as references for annotation. Although it seems unlikely that the genomes of all the 11 million extant worldwide species will ever be or need to be sequenced, the genomic sequences for a large number of eukaryote species are already available for scientific scrutiny, including the genomes of some endangered vertebrate species that may need assistance in the management of their breeding and survival [122]. In 2009, an international consortium established the Genome 10K Project to sequence and analyze the complete genomes of 10,000 vertebrate species (http://genome10k.org).

NGS has blasted human genomics into an exciting new era of genetic investigation geared towards humanomics and disease (see section *8.9*) and the management of an individual's life cycle and health issues by way of personal genomes or personomics [123]. Targeted or whole-genome resequencing of individuals from within the same or different species is aimed to detect and catalogue SNPs, mutations, and sequence variants such as indels, copy number,

and structural variations [14–16]. PCR-based candidate gene and whole-exome analysis are two widely used methods that can be performed with higher coverage and at much lower cost than whole-genome resequencing. Genotyping HLA genes of humans for clinical diagnosis or research by sequencing the entire gene [97, 128] or just the exons [129] is an example of targeted resequencing to identify polymorphisms that are important in tissue or cell matching for transplantation [130]. Exomics is targeted specifically towards coding genes and discovering exonic mutations responsible for rare Mendelian disorders such as hearing loss, intellectual disabilities, and movement disorders and for investigating common disorders such as heart disease, hypertension, diabetes, and cancer [13, 123, 125], and many others that are listed at the Online Mendelian Inheritance in Man (OMIM) database (Table 3, [49]). In contrast to WES, WGS can assess alterations in the coding genes and the regulatory and noncoding regions [123, 126], especially multiallelic copy number variations [127]. Cancer research has shown that it is important to target all types of somatic/germ-line genetic alterations, including nucleotide substitutions, small insertions and deletions (indels), CNVs, and chromosomal rearrangements in the noncoding regions [13, 15, 123]. WGS has been used to identify variants, indels, and multiple numbers of genes involved in rare and common diseases such as Charcot-Marie-Tooth neuropathy, dopa-responsive dystonia, acquired essential thrombocytosis, erythrocytosis, and many others [123, 126].

8.2. Transcriptomics and RNA sequencing

RNA-seq is the NGS method that sequences the transcriptome, that is, all the RNA transcript sets expressed by the genome in cells, tissues, and organs at different stages of an organism's life cycle [12, 18, 19, 20, 30]. High-throughput RNA sequencing using cDNA fragments was first employed in mammalian cells [131] and yeast [132], and now it is used for a wide range of organisms [133]. Without transcriptome data, the genome sequence alone is of limited use for understanding the intricacies of genome function in biology. RNA-seq provides technical reliability and sensitivity and unambiguous maps of the transcribed regions of the genome with high accuracy in quantitative expression levels, identification of tissue-specific transcript variants and isoforms (SNPs and mutations), transcription boundaries and splicing events, transcription factors, and small and large noncoding RNAs (ncRNA) involved in the regulation of gene expression [131–137].

At least 90% of the mammalian genome is actively transcribed to produce different classes of ncRNAs [135, 136], including ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), small interfering RNA (siRNA), PIWI-interacting RNA (piRNA), and large intergenic noncoding RNA (lincRNA) [138–141] and retrotransposons [142–146]. The known classes of functional ncRNAs consists largely of those supporting protein translation (ribosomal, transfer, and small nucleolar RNAs), transcript splicing (snRNAs) [137, 138], and miRNA that target conserved binding sites of mRNAs to decrease their stability [139]. The new class of small piRNA was discovered to interact with PIWI regulatory proteins and RNA to silence transposons in the germ line and regulate gene expression in the soma [140]. The lincRNAs are expressed by a different class of actively transcribed RNA genes and they have diverse roles in processes such

http://dx.doi.org/10.5772/61964

as cell cycle regulation, immune responses, brain processes, and gametogenesis [147–150]. A substantial fraction of lincRNAs binds to chromatin-modifying proteins and may modulate gene expression by bringing together protein complexes for specific functions [150].

Defective splicing of transcripts and expression levels are believed to contribute to at least 50% of inherited human diseases [151]. Altered expression levels of specific isoforms or alleles have been identified in ischemic stroke, type 2 diabetes, colorectal cancer, chronic lymphocytic leukemia, and many other diseases [30]. Dysregulation of gene expression, splicing, and other editing events in specific cell types have been associated also with the pathogenesis of cardiovascular diseases, neurological disorders, and different cancers [137, 151–153]. Similarly, different classes of small and large ncRNAs have been found to be associated with different diseases and cancers [147–149]. The expressed information of the transcriptome varies enormously between different cells of a multicellular organism and depends on the cell type and its functional and temporal state. At least two important databases, the Encyclopedia of DNA Elements (ENCODE) and Genotype-Tissue Expression (GTEx) (Table 3), have focused on mapping functional elements at high resolution and the regulation of gene expression and the transcriptome in different tissues of humans. The GTEx project is one of the most recent projects that have generated a large amount of RNA sequence data by RNA-seq technology to investigate the patterns of transcriptome variation across 43 tissues and 1641 samples from 175 postmortem individuals [153]. The analysis included 20,110 protein-coding genes and 11,790 lncRNAs with 88% and 71%, respectively, detected in at least one sample. A relatively small number of genes (a few hundred) were expressed for most tissues with a definite, differential modular profile showing tissue-preferential expression. In addition, 3,046 protein-coding genes were expressed together with an adjoining repeat element such as Alu, L1, ERV, Tigger, and Charlie [153]. These findings provide a better systematic understanding of the heterogeneity among a diverse set of human tissues and the enormous complexity and variation involved in the regulation of genome expression.

8.3. Methylomics and epigenomics

Epigenomics is the study of heritable gene regulation that does not involve the DNA nucleotide coding sequence itself but acts on a genome-wide scale via DNA nucleotide methylation and posttranslational modifications of histones, the interaction between transcription factors and their targets, and nucleosome positioning [23–30]. Methylomics is the genome-wide analysis of DNA methylations and their effects on gene expression and heredity [28]. Methyl-seq uses NGS to analyze and map DNA cytosine methylation at single-base resolution usually by employing bisulfite DNA sequencing [24, 25]. Treatment of genomic DNA with sodium bisulfite converts cytosines but not methylcytosines to uracils so that the uracils are PCR converted and sequence differentiated at the SNP locations as thymidines and the methylcytosines are sequenced as cytosines. Bisulfite DNA sequencing is used widely for DNA methylation profiling in various organisms as well as humans for assessing disease genes [23, 27, 29].

ChIP-seq is chromatin immunoprecipitation (ChIP) that is followed by NGS sequencing. It permits genome-wide profiling of DNA-binding proteins and histone and nucleosome

modifications [30]. The ChIP-seq technology was partly adapted from microarray ChIP-chip technology and first implemented in 2007 and since then has been used widely to analyze transcription factor binding sites, histone modifications, and chromatin-modifying complexes and sequences in a wide variety of organisms [154]. ChIP-seq provides higher resolution, less noise, and greater coverage than the array-based ChIP-chip method that was previous used, and therefore, it has become the preferred tool for studying gene regulation and epigenetic mechanisms. Two other NGS tools commonly used for epigenetic studies are Hi-C and ChIA-PET that provide insights into the global 3D organization of eukaryote genomes [30]. Hi-C utilizes NGS on cross-linked DNA fragments to identify the DNA regions such as promoters, enhancers, and insulators that come together to mediate their regulatory activities. ChIA-PET uses immunoprecipitation of crosslinked-interacting protein-DNA and paired-end sequencing to reveal the interaction between enhancer and promoter regions located at intergenic distances away from each other but either on the same (*cis*) or different (*trans*) chromosomes [30]. de Wit and de Laat [155] provided an overview of the various derived chromosomal conformation capture (3C) methods, including 4C (chromosome conformation capture-on-chip) and 5C (chromosome conformation carbon copy) and their application in the study of chromatin interactions. Two epigenomic databases on the Internet, the NIH Roadmap Epigenomics Project and Blueprint (Table 3), catalogue the chemical modifications to the genome and how they activate gene expression in human tissues and cell types.

8.4. Proteomics, metabolomics, and systeomics

Proteomics is the large-scale study of the structure, function, identification, and characterization of peptides and proteins [113, 156, 157]. This includes information on protein abundance, variations and polymorphisms, modifications, and their interactions and networks in cellular processes. As a first step, the sequence translation of open reading frames of genomes, exons, and transcripts using a codon table and one or more bioinformatics tools is the simplest way of constructing proteomic profiles from NGS data. However, this is not the only analytical protocol used in the domain of proteomics, and protein specialists often employ a variety of other hardware and software tools to build up an organism's peptide and protein profiles. Among these are the detection and analysis of proteins and their functions by two-dimensional polyacrylamide gels, liquid chromatography coupled with tandem mass spectrometry, affinity-tagged proteins, and yeast two-hybrid assays [156, 157]. A number of public databases for proteomics and protein-protein interactions are available on the Internet such as ExPASy and PRIDE (Table 3).

Metabolomics is the study of an organism's total metabolic response to an environmental stimulus or a genetic modification [113]. The metabolomics of organisms are drawn indirectly from NGS data, mainly from the known functions of enzymes and proteins involved in metabolic and biochemical pathways. Metabolomics data also provide biochemical and physiological snapshots of processes that are obtained from cellular and tissue experimental studies using various technologies of separation (gas chromatography, high-performance liquid chromatography, and capillary electrophoresis) and detection (mass spectrophotometry, NMR spectrometry, ion mobility, and thin-layer chromatography) [158]. Metabolomics is

an important part of functional genomics for determining the phenotypic effects of genetic manipulations such as gene deletions, insertions, and mutations. Nutrigenomics is an arm of metabolomics that links genomics, transcriptomics, proteomics, metabolomics, and microbiomics together in an examination of the effects of nutrition and energy metabolism on gene expression in relation to an organism's genotype [113, 159]. The use of constraint-based methods such as the Flux Balance Analysis to design models of metabolite flow in microbes has connected "omic" to phenotypes in the science of Fluxomics [160]. Some web-based metabolomic resources include FAME, AromaDeg, Metabolomexpress, and MetaboAnalyst (Table 3).

Systeomics is the integration of genomics, proteomics, metabolomics, and phenomics into a single network system. It is a branch of biology that uses computational techniques to analyze and model how the components of a biological system such as cells or organisms interact with each other to produce the characteristics and behavior of that system [160–162]. Systeomics is a biology-based interdisciplinary field applied to biomedical and biological scientific research that focuses on complex interactions within biological systems using a holistic approach. For example, the U.S. Department of Energy's Genomic Science program uses microbial and plant genomic data, high-throughput analytical technologies, and modeling and simulation to develop a predictive understanding of biological systems behavior relevant to solving energy and environmental challenges (http://doegenomestolife.org). The U.S. Department of Energy Systems Biology Knowledgebase (KBase) is a software and data platform for systems biology mechanisms (Table 3) to assist with the prediction and design of biological functions of microbes and plants. KBase integrates data, tools, and their associated interfaces into one unified, scalable environment to allow users to upload their own data for analysis, to build models, and to share and publish their workflows and conclusions. Another example is the Kyoto Encylopedia of Genes and Genomes (KEGG), which is a database resource to integrate high-level functions and utilities of biological systems from molecular-level information (Table 3). Other "omics" that contribute to the "omic" lexicon and biology are epidemiomics [163], physionomics [113], variomics [164], and phenomics [165–167]. In the case of phenomics, the European Genome-phenome Archive (EGA) provides accession numbers that refer to the relationship between genomics and phenotype/traits, such as the physical and biochemical traits of humans (Table 3). It integrates physical traits or phenotypes with genomics, transcriptomics, methylomics, proteomics, and metabolomics [166].

8.5. Metagenomics and microbiomes

Metagenomics, or beyond genomics, is the study of the total genomic content of a microbial community that bridges the three domains of life, Archaea, Bacteria, and Eukaryotes [100, 114–118, 168–179]. The total DNA and/or RNA is isolated from a microbial population without prior cultivation, sequenced, and compared with previously known sequences to identify known species or to discover previously unknown species. Some of the environments from which microbial communities are isolated and studied include aquatic and terrestrial environments, host-associated ecosystems, and various human engineered systems such as those involved with food, water, and waste production, agriculture, animal husbandry, and various

human and animal habitations [100, 115, 168, 169]. Hospitals are a worrying source of pathogenic microorganisms, especially those that develop resistance to commonly used medical antibiotics [115, 168]. Thus, NGS is an important growing application for epidemiological studies of various pathogens, such as mycobacteria, *S. aureus*, *E. coli*, cholera, influenza, HIV, Ebola virus, etc. [169–171]. The Earth Microbiome Project (http://www.earthmicrobiome.org) is an ambitious multidisciplinary attempt to analyze microbial communities across the globe using approximately 500,000 reconstructed microbial genomes.

The earliest metagenomic studies targeted 16S rRNA genes to genotype and identify the different species within the environment before the first NGS microbial studies using the Roche pyrosequencing and Illumina platforms targeted mining sites and the surface waters of the gulfs, seas, and oceans [114, 169]. Many big projects and consortia for sequencing metagenomes have been launched in the past 10 years, such as the TerraGenome project for soils (Table 3) and the *Tara* Oceans project on the microbiome, eukaryotic plankton, and viromes of the global oceans [172–174].

Microbes colonize the human body (microbiome) in numbers that are estimated to outnumber human genes and somatic cells by more than 100-fold [175]. These microbes (viruses, prokaryotes, and eukaryotic microbes) occupy various anatomical habitats including gut, skin, vagina, and oral mucosa and are believed to markedly influence human physiology, nutrition, and health [175–177]. Advances in NGS and computing methods have enabled human microbiome studies such as the MetaHit project and the Human Microbiome Project (HMP) (Table 3). In May 2015, SRA that was established by NCBI in 2008 to store raw sequence data from the NGS technologies had over 2,068 trillion open access nucleotides in its database to massively outgrow GenBank, EMBL, and DDBJ for bacterial sequence storage. The genomic sequences continue to accumulate in other databases as well [114], such as 47,083 prokaryotic genomes projected for Genomes Online Database (GOLD) [178] and 152,927 metagenomes for the MG-RAST server [179]. As of October 2014, the GOLD database contained 544 metagenomics studies associated with 6726 metagenome samples, whereas MG-RAST held 150,039 metagenomic samples, of which 20,415 were publicly available (Table 3). Recently, Zelezniak et al. [180] gathered and modeled NGS 16S rRNA sequence data to map interspecies metabolic exchanges and resource competition based on the genomic potential encoded by the microbial communities. They analyzed more than 1,297 communities and 261 species in soil, water, and human gut samples and concluded that the interplay between competitive and cooperative strategies for resources and the ability to exchange metabolites, such as amino acids and sugars, shapes the composition of microbial communities.

8.6. Comparative genomics, phylogenomics, and the phylomes of life

Comparative genomics and phylogenomics via NGS and the phylome (complete collection of all gene phylogenies in a genome) provide powerful applications for classifying and understanding the differences and similarities of all life forms and for unraveling their evolutionary histories [100, 116, 176, 181–186]. The three basic domains of life, Bacteria, Archaea, and Eukarya, were first identified and classified phylogenetically on the basis of ribosomal RNA sequences [181]. Although Bacteria and Archaea are both placed into the

kingdom of the Prokaryotes or Monera (lacking a membrane-bound nucleus, mitochondria, and chloroplasts but containing a cell wall), their separate rRNA sequence clusters clearly divided them into distinct domains [181]. The Eukarya (eukaryotes) have been subdivided into four basic kingdoms, Protista, Fungi (Mycota), Plantae (Metaphyta), and Animalia (Metazoa) [182]. However, on the basis of metagenomic and phylogenomic studies and NGS data, the classifications and nomenclatures of eukaryotes continue to be revised and organized into other supergroups such as Amoebozoa, Opsthokonta, Ecavata, Archaeplastida (Plantae), SAR (Stra/Alveo/Rhizaria), and Incertae sedis [183, 184]. On the other hand, because viruses do not have rRNA genes, they have missed out on a life-domain classification [185, 186]. There is still a strong debate about whether or not viruses without rRNA genes should be classified as a separate life form (a fourth domain) or simply be viewed as exogenous parasites, infectious agents, and endogenous mobile elements that are dependent on and exist within the life forms of the three defined domains [185, 186]. Viruses impact greatly on all life forms, so they are a major interest for NGS applications and phylogenomics [34, 114, 174, 187–189], especially emerging viruses such as dengue, Ebola, Chikungunya, MERS, lyssavirus, and norovirus (http://viralzone.expasy.org), which are of a great concern to human health [114, 171, 189].

NGS, phylogenomics, and taxonomy profiling during the past decade has greatly expanded our knowledge of the diversity of bacterial genomes from the same and different species [116, 190], with the discovery of many DNA sequence repeats and transposons that contribute to at least 10% of the genome and play an important role in immunity [100, 191]. Archaea and thermophiles have a large proportion of their genomes consisting of defense genes often localized in genomic islands as a consequence of horizontal gene transfers [191, 192]. For example, the family of clustered regularly interspaced short palindromic repeats (CRISPRs) and the CRISPR-associated proteins in the CRISPR-Associated System (CAS) that have an important role in the host's adaptive immunity to pathogens and as responders to environmental stress [192–194] have been translocated between different prokaryote strains and species [191, 192]. CAS includes distinct gene families of 50 or more that show strong evidence of extensive plasticity and horizontal gene transfer to protect prokaryote cells against the replication of phage and plasmids that integrate into the CRISPR locus [193–195]. Moreover, the CRISPR/CAS systems have been developed as an "in vitro" genetic engineering tool to be transfected into the cells of various organisms to manipulate their genes [196], including the foreign defense system introduced into human cells against HIV-1 infection [197]. Other bacterial defense systems that have been studied or discovered by NGS include the toxin/anti-toxin, antigen, novel restriction-modification, and DNA phosphorathioation systems as well as those involved with infection-induced dormancy or programmed cell death [192]. Genomic sequencing also has revealed new bacterial microcompartments, protein structures, or organelles that are used in metabolic pathways [198], such as those involved in carbon fixation and metabolism of amino alcohols, ethanol, rhamnose, and fucose [199]. Bacterial genomes also provide sequences for phylogenetic and gene comparisons, taxonomic classification, transcriptomics, and methylomics and for the assessment of sequence diversity and variants for a better understanding of gene func-

tions [100]. Although the classical operon structure predominates in bacteria and archaea, a variety of other transcription unit architectures have been elucidated [100]. More than 4,661 transcription units have been described with an average of 1.7 promoters per operon, and transcription factor binding sites have been determined for virtually all the transcription factors in *E. coli* [100]. DNA methylation was first discovered in bacterial restriction-modification systems with diverse functions in addition to cellular defense [200], and it is now seen as an evolutionarily conserved form of transcriptional repression and an ancestral form of defense against foreign DNA molecules and transposons and other mobile elements in all life forms [201].

Phylogenomics has been used to reevaluate the evolutionary affiliation between archaea and eukaryotes and to infer that the nuclear lineage in eukaryotes emerged from the archaeal radiation and most probably from the archaeal TACK superphylum [202]. Recently, Spang et al. [203] sequenced uncultivated metagenomes from a deep-sea vent and discovered novel archaeal genomes in the new phylum that they named "Lokiarchaeota." These novel archaea contain homologues of many eukaryotic proteins that function in the endomembrane system and in phagocytosis, including actin and related proteins, and Ras superfamily GTPases, suggesting that this newly discovered phylum is the missing link in eukaryogenesis. Although eukaryotes possess the membrane-enclosed mitochondrial organelle and prokaryotes do not, the eukaryotic mitochondria are believed to have evolved from a bacterial system, probably by endosymbiosis [204] involving an ancestor within the bacterial phylum Alphaproteobacteria [205]. Although mitochondrial phylogenomics suggests a monophyletic origin and assemblage, it is now evident that the mitochondria are genetic chimeras and functional mosaics with the bulk of the mitochondrial proteome originating during eukaryote evolution outside the Alphaproteobacteria and other bacterial phyla. It seems that the mitochondrial genome has expanded and contracted in various lineages during evolution with much of the original mitochondrial genetic information transferred to the nucleus [205]. Eukaryotic diploid cells appear to have evolved 2 billion years after haploid prokaryotes, and their evolution from proto-eukaryotic cells, such as the multinucleated *Giardia* organism [206], seems to have involved chromosomal crossing over from mitotic recombination to meiosis and to sexual reproduction where a set of chromosomes is inherited from each parent [207]. The genomes of diploid eukaryotes are usually larger than those in haploid prokaryotes probably because greater information complexity is needed by multicellular organisms to regulate and coordinate the multiple stages of their life cycles with the added requirement for more molecular regulatory systems to communicate and interact between multiple tissues and organs [206].

Eukaryotic genomes vary markedly in size and gene number and appear to be variable in their susceptibility to polyploidy (a doubling of the diploid sets of chromosomes), redundancy, duplication, and the persistent accumulation of interspersed repeats and mobile elements [208–210]. For example, the genomes of plants can range from the simplest like *Ostreococcus tauri* with a 12.6 Mb genome, containing less than 8,000 genes and minimal genome duplication [211], to the highly complex such as the canopy and pale-petal flowering plant *Paris japonica,* with a 150 Gb genome and eight sets of chromosomes derived by allopolyploi-

http://dx.doi.org/10.5772/61964

dy and hybridization of four species [212]. The genomic size of *Paris japonica*, which has still to be fully sequenced, is 50 times larger than the human genome and extends the range of genome sizes to 2,400-fold across angiosperms and 66,000-fold across eukaryotes [212]. Genome duplication and polyploidy, both recent and ancient, have contributed to the considerable genomic complexity in eukaryotes, particularly in plants, amoeba, fungi, and vertebrates [208–223]. Following ancient polyploidization, most duplicated genes are deleted by intrachromosomal recombination, a process referred to as fractionation, and any remaining evidence for the polyploidy event is not easy to find by phylogenomic analysis [214]. Nevertheless, a phylogenomic comparison of gene duplications in a four-way comparison of paralogous regions in tunicate, fish, mouse, and human provided unmistakable evidence of two distinct genome duplication events (the 2R event) early in vertebrate evolution and before the divergence of fish and mammalian lineages [215], as was proposed by Ohno in 1970 [216]. Interestingly, polyploidy also can occur in humans during normal development and cancer [208, 209]. Fetal polyploidy in the form of triploidy (69,XXX chromosomes) and tetraploidy (92,XXXX chromosomes) is a rare and lethal event, resulting in spontaneous abortions or brief postpartum survival times [208], whereas polyploidy is common in stressed tissues and cells and in tumor development [208, 209]. On the other hand, comparative genomic studies have revealed that polyploidy is common in the evolutionary history of many different flowering plants [208, 214], for example, between different species of the allopolyploid tobacco plants, *Nicotiana* section Repandae [217]. In comparing the allotetraploid genomes of *Nicotiana repanda* and *Nicotiana nudicaulis,* it was assessed that the loss of low-copy sequences along with the loss of duplicate copies of genes and upstream regulators reflects genome diploidization, whereas genome size divergence between the allopolyploids is manifested through differential accumulation and/or deletion of high-copy-number sequences and transposable elements [217]. Diploidization and genome size change in *Nicotiana* allopolyploids is associated with differential dynamics of low- and high-copy sequences [218]. The induction of polyploidy is a common technique to overcome the sterility of a hybrid species during plant breeding; therefore, many agriculturally important plants such as the genus *Brassica* are polyploids [219-221]. Wheat, after millennia of hybridization and modification by humans, has strains that are diploid (2 sets of chromosomes), tetraploid (4 sets of chromosomes), and hexaploid (6 sets of chromosomes) [222, 223], whereas the invasive weed *Spartina anglica* has up to 12 sets of chromosomes [224].

A recent comparative genomic study has revealed how genomes change with speciation in an examination of genomes from five cichlid fish species, an ancestral lineage from the Nile, and four species from the East Africa lakes, Tanganyika, Malawi, and Victoria [225]. Compared to the ancestral Nile lineage, the East African cichlid genomes had many alterations in regulatory elements, accelerated evolution of protein-coding elements in genes for pigmentation, an excess of gene duplications, and other distinct features that affect gene expression associated with transposable element insertions and novel microRNA. Each species contains a reservoir of mutations different from the other species [225]. Much of the diversity between species evolves in a nonparallel manner often rapidly due to sexual selection and genetic conflicts

between males and females and between different regions of the genome at a regulatory level rather than by the slower and weaker forces of classical natural selection [226].

Most genomes range between newly derived genes and the ultraconserved or the essential core coding and noncoding genes [100, 227, 228]. Comparative genomics has resulted in the discovery of ultraconserved noncoding elements (UCNE) across different phyla, starting with 481-long segments (>200 bp) that are 100% conserved between orthologous regions of the human, rat, and mouse genomes and 95% to 99% conserved in chicken and dog genomes [229]. A more recent comparison of 28 vertebrate genomes identified millions of additional conserved elements with distinct types of functional elements including regulatory motifs present in the promoters and untranslated regions of coregulated genes, insulators that constrain domains of gene expression, and conserved secondary structures in RNAs and in developmental regulators [230]. A webpage at http://ultraconserved.org provides study protocols, computer software, and references dedicated to ultraconserved elements [229]. Also, there are at least two databases for the conserved noncoding elements and the genomic regulatory blocks (Table 3), the UCNEbase for human and chicken [231], and the UCbase 2.0 for the 481 UCNE that were longer than 200 bp and that were discovered in the genomes of mammals [229]. The UCNEbase suggests that the evolution of species depends more on innovation and change in regulatory sequences than in proteins [231]. Indeed, there are essential genes that are indispensable for the survival of an organism and therefore are considered a foundation of life. The database of essential genes (DEG) (Table 3) catalogues known essential genomic elements, such as protein-coding genes and noncoding RNAs, within the bacteria, archaea, and eukaryotes that constitute a minimal genome and are useful for annotating newly sequenced genomes [232].

Phylomes provide the combined analysis of genome-wide collections of phylogenetic trees to aid in the inference of orthological and paralogical relationships and the detection of evolutionary events such as whole-genome duplication (polyploidization), gene family expansion and contraction, horizontal gene transfer, recombination, inversion, and incomplete lineage sorting [233, 234]. The online PhylomeDB v4 database was created as a phylogenomic repository and is useful for preliminary phylogenetic data analysis of genomes of interest from various phyla as well as for annotating newly derived genomic sequences [234]. As an example, Fig. 1 shows the PhylomeDB analysis of the duplications of the RLTPR gene, a gene that was first discovered in humans in 2004 [235]. The PhylomeDB analysis shows that the RLTPR gene has two paralogs, LRRC16B and LRRC16, which were generated by two separate duplication events at least prior to the divergence of mice and humans (Fig. 1). The functions of RLTPR are not well characterized, but its distinct functional domains suggest that it may multitask in protein-protein interactions, as recently demonstrated in the development of regulatory T cells in mice [236]. The analytical approach to find orthologous and paralogous relationships with maximum genomic coverage for the RLTPR gene is both gene-centric and genome-wide in PhylomeDB. Also of particular interest are the well-conserved genomic mechanisms of innate immunity, such as Apolipoprotein B Editing Catalytic subunit proteins 3 (APOBEC3s) in mammals that mutate and inactivate viral genomes [237]. Other phylogenetic databases that complement PhylomeDB in a comparative analysis are Ensembl Compara GeneTrees, TreeFam, PANTHER, PhyloFacts FATCAT, and the HOGENOM database (Table 3).

http://dx.doi.org/10.5772/61964

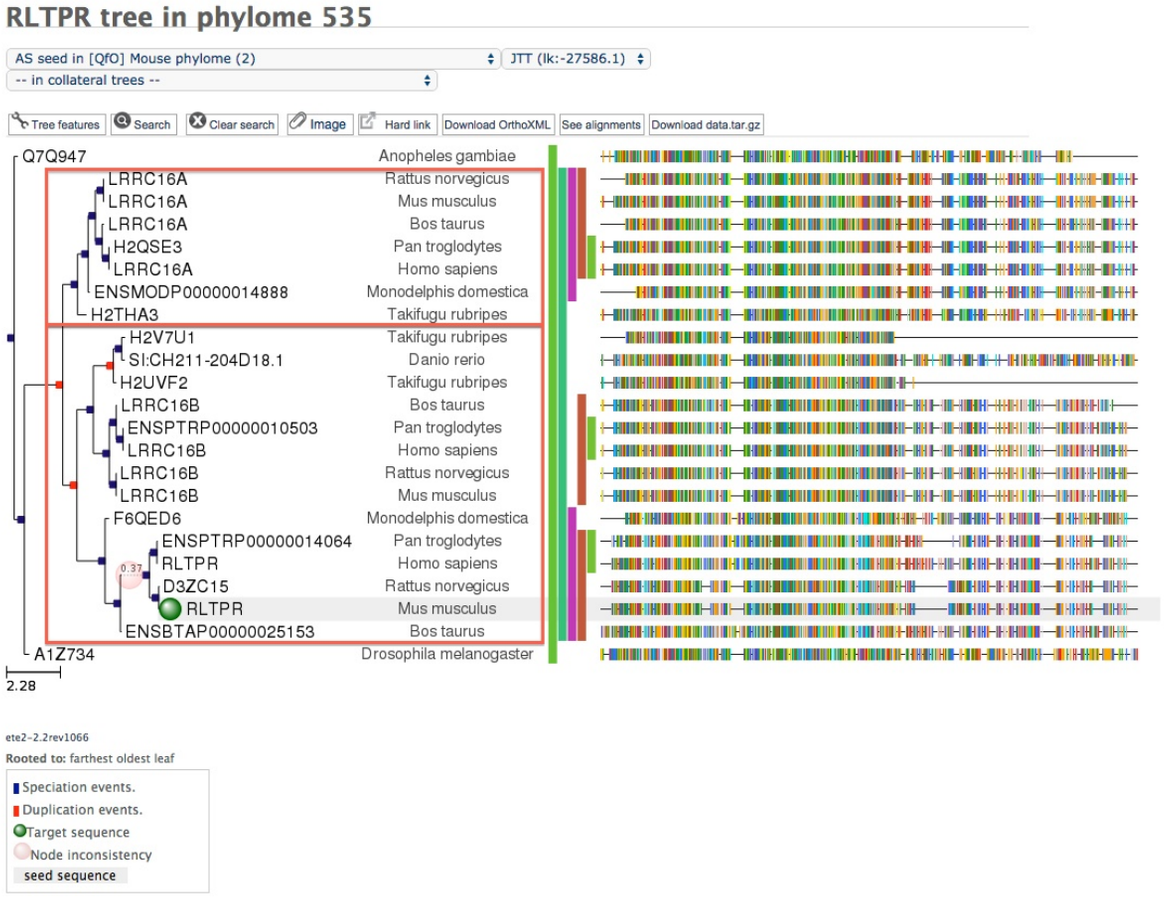

Figure 1. RLTPR gene tree shows the RLTPR gene orthologs and paralogs in 10 vertebrate species. The human gene RLTPR (NCBI Gene ID: 146206), first reported in 2004 [235], was used as the search query for the Phylome tree at http://phylomedb.org with the phylome data settings of AS seed in (Qf0) mouse phylome (2) and JTT (lk:-27586.1). The tree shows the speciation events (blue squares) and three duplication events (red squares) at the nodes with the first duplication event arising early in vertebrate evolution before the divergence of fish and mammalian lineages [215].

8.7. Mobilomics and Horizontal Gene Transfer (HGT)

The science of mobile genetic elements (mobilomics) developed long before the advent of genomics and NGS [238]. The 1983 Nobel Prize winner Barbara McClintock first reported the existence of mobile elements as jumping genes in maize in the late 1940s [239]. The discovery of new classes and families of DNA transposons and autonomous and nonautonomous retrotransposons continued slowly for the next five decades until the first online repeat element screening webserver CENSOR and database REPBASE (Table 3) was established by Jerzy Jurka and his colleagues between 1992 and 1996 [240, 241]. Since then, RepeatMasker (Table 3) and other tools such as Mobster [242], Red [243], and Visual TE [244] have followed on to help define the mobilome, the totality of mobile genetic elements in a particular genome. A list and description of some of the families, types, and classes of transposons and retrotransposons in prokaryotes and eukaryotes can be found in the following reviews [238, 245–251]. A recent survey of repeats and mobile elements that affect genomic stability has elucidated how some bacteria can control the mobilome through postsegregation killing systems

[192–195, 247]. Different classes of TEs are found in the genomes of different eukaryotes that contribute to at least 50% of the human genome [237] and up to 90% of the maize genome [252]. In humans, there are solitary Long Terminal Repeats (LTR) and LTR retrotransposons (endogenous retroviruses) that are characterized by the presence of LTR at both ends; Long Interspersed Nuclear Elements (LINEs) like L1 that represent families of non-LTR TEs about 6 kb in length and encode two proteins, a nucleic acid chaperone, and a reverse transcriptase/nuclease for retrotransposition; Nonautomomous Miniature Inverted-Repeat Transposable Elements (MITEs); Mammalian-wide Interspersed Repeats (MIRs), an ancient family of tRNA-derived SINEs exapted as enhancers and regulatory sequences; and Short Interspersed Nuclear Elements (SINEs) like Alu that are usually less than 300 bp and need a helper transposon element like L1 for transposition [245]. Most ERVs, SINEs, and LINEs in the human genome are now remnants of past insertions and are no longer capable of actively ”jumping” like functional TEs [238, 245, 248]. Indeed, many of the TE ancient relics have undergone exaptation and developed new functions, such as transcript repeat elements, within regulatory gene networks to generate lineage-specific adaptation [145, 249].

The importance of widespread HGT in creating genomic diversity in microbes has been highlighted by the many comparative genomic studies using metagenome data [191]. Comparative genomic analysis of different strains of *E. coli* revealed that up to 30% of genes in pathogenic strains were acquired by HGT often creating duplication events and modifying metabolic networks by adding operons that encode two or more enzymes [253]. Comparative genomics of photosynthetic prokaryotes revealed that they have evolved as complex mosaics via multiple HGT events [254]. Similarly, photosynthetic gene clusters and gene clusters that encode various toxins, resistance genes, metabolic genes, and components of secretion systems appear to be the products of HGT [247, 253–255]. Indeed, many HGT events probably were mediated by genomic mobile elements, such as bacteriophages, plasmids, viruses, transposable elements, and toxin/antitoxin systems that are persistent in all life forms [191, 228, 246, 255, 256].

Before the new millennium, transposons and repeat elements were largely viewed as junk and as parasites that created unnecessary burden on the genome. Comparative genomics and online databases dedicated to transposons and repeat elements such as SINES, LINES, and ERVs, however, began to change this picture in the 1990s, and it soon became evident that these elements were the drivers of evolutionary innovation. Many integrated transposons mutate with time to interact with the host transcriptional machinery and therefore provide a useful substrate for evolution of novel regulatory elements [145, 228, 255–258]. Moreover, some of the ancient integrated retrotransposons appear to have been involved in advantageous segmental genomic duplications such as in the major histocompatibility complex region [259–261], and others have dispersed regulatory controls to provide coordinated regulation across the genome [257, 258].

8.8. Agrigenomics

Agrigenomics or agricultural genomics can be defined as the research and development activities that translate NGS and genomics technology into a better understanding of plant

biology and advancing crop improvements. During the past decade, NGS had an enormous impact on developing fundamental genome resources to directly address many of today's concerns in agriculture and agronomics. Since the publication in 2000 of the first plant genome, *Arabidopsis thaliana*, 54 new plant genomes were published by 2013 [221] followed by at least another 6 plant genomes including the hexaploid bread wheat genome [223]. In reviewing the first 55 plant genomes, Michael and Jackson [221] concluded that, although these genomes have provided a glimpse at the gene number, types, and numbers of repeats and genomic growth, contraction, and rearrangement, we are only just at the beginning of defining the functional aspects of plant genomes "and various other 'omics' data layered on genomes."

8.9. Humanomics, personomics, and health

The accumulation of knowledge on the human genome and its genetic and molecular processes (humanomics) has amplified considerably since the first draft assembly was published in 2001 [262]. The first human hybrid genome took about 15 years to sequence and assemble, and when released to the public, it covered 90% of the euchromatic genome, contained about 250,000 gaps, and had many errors in the nucleotide sequence [43, 44]. Ten years after the publication of the first human draft sequence, six more human genome sequences were completed with a much greater coverage and accuracy, enabling more informative comparisons to be made between them [7, 8, 79]. Studies by the 1000 Genomes Project [10], the Personal Genome Project [263], the HapMap Consortium [264], and the Pan-Asian Single Nucleotide Polymorphism Project [265] revealed the enormous sequence diversity that exists between individuals. Since then, 225 Ethiopian and Egyptian genomes were compared to reconstruct their population history out of Africa [266], 911 genomes from 10 populations of African, East Asian, and European ancestries were sequenced to elucidate novel patterns and signatures of genetic differentiation [267], and whole-exome sequences from 951 genomes of a ClinSeq cohort were compared to discover new loss-of-function mutations [268]. Today, there are many 1000 human genome projects, and WGS of the human genome for personalized medicine (personomics) is already a reality for 2,638 Icelanders [9] and for some others [269, 270] of the 7.3 billion individuals currently populating the globe (http://www.worldometers.info/world-population).

Veeramah and Hammer [271] recently reviewed the usefulness of NGS to sequence ancient DNA samples for phylogenetic and evolutionary studies and for the reconstruction of human population history. Some of these NGS studies have helped to refine the demographic histories of human evolution. These studies include those of the ancient DNA of extinct hominins (Neanderthals, Denisovans) and ancient modern humans such as 7,000-year-old Mesolithic hunter-gathers in northwestern Spain, Neolithic and post-Neolithic (5,300- to 4,000-year-old) hunter-gathers and farmers in Scandinavia, a 4,000-year-old Paleo-Eskimo from southern Greenland, and a 24,000- and 17,000-year-old South-Central Siberian [271]. NGS of ancient nonhuman genomes such as those of pathogens, parasites, and domesticated animals and plants also can provide new information about human history in regard to life styles, health, and the spread of agriculture [272].

NGS has allowed a detailed analysis of single nucleotide variants (SNVs), structural variants (SV), and methylations in coding and noncoding regions and to assess their role in human disease [9, 14, 15, 19, 22, 25, 29, 30, 123, 125–127, 144, 148, 151, 152]. The establishment of the International HapMap Project in 2003 (Table 3) to develop a "hapmap" of human haplotype genomes from samples of large populations was an important initiative to find genes and genomic variations (SNP and CNV frequencies, genotypes, and phased haplotypes) that affect health and disease [264]. More than 97 million validated SNPs (dbSNP) have been discovered from human genome sequencing projects and many of the variants have been linked to a range of medical and phenotypic conditions and catalogued at dbGAP (Table 3), the database of genotype and phenotype [273]. In July 2015, dbGAP had links to 592 disease and phenotype studies and 3,711 data sets. In addition to SNV, small and large SVs that are duplicated, deleted, or rearranged relative to the reference sequences and individuals have been identified in NGS studies and associated with various diseases [9, 30, 127]. NGS has been used to diagnose rare Mendelian diseases and genetically heterogeneous complex disorders, such as X-linked intellectual disability, congenital disorders, cancer genome heterogeneity, and fetal aneuploidy [13, 15, 123, 125, 208, 209, 274, 275]. The impact of NGS on the diagnosis of rare genetic diseases is evidenced by the growth of the genes and OMIM database [49, 276] that has doubled in data since 2007 [274]. However, it should be noted that NGS does not always reveal causative mutations but instead may provide a list of possible candidates. Many detected SNPs, SNVs, and SV have not been associated to disease or phenotype and many diseases still await a genetic or genomic cause. NGS in human studies must be used with caution because of the significant levels of false-positive and false-negative rates in sequencing errors and amplification biases.

Soon et al. [30] listed and reviewed the various NGS methods employed in the ENCODE project to annotate and analyze the transcriptome and map elements and identify the methylation patterns of the whole human genome. The information in ENCODE and other databases such as GTEx, FANTOM, NIH ROADMAP, and BLUEPRINT (Table 3) has enabled researchers to map genetic variants to gene regulatory regions and assess indirect links to disease. The Regulome DB based on the accumulation of nongenic functional regulatory regions obtained from ENCODE is a useful resource for the evaluation of polymorphisms of regulatory regions [276]. Although disease-associated SNPs obtained from GWAS studies may point to gene coding regions, they actually might reside in regulatory sites of downstream genes that are in linkage disequilibrium with the reported SNPs [262]. RNA-seq and NGS has confirmed that 98% of the human genome is transcribed from noncoding genomic regions, that only about 2% of the human genome codes for peptides and proteins with about 20,000 distinct protein-coding genes, and that alternative splicing seems to occur for 90% of protein-coding genes to yield many more different types of proteins than genes [134, 135, 151, 152]. The vast majority of the human genome is not functionless "junk DNA" as previously thought [262], but rather, it can be viewed as DNA/RNA "dark matter" expressing hundreds to millions of transcribed short and long noncoding RNA molecules that have important regulatory roles in transcription, translation, transport, metabolism, and innate immunity [133]. Some of these are the interspersed retroelements such as Alu and L1 and endogenous retroviruses (ERVs) that have evolved before and during primate history to function as regulators of transcription and translation [24, 25, 142, 145, 257, 258].

NGS has especially revolutionized the field of cancer genomes revealing mutations, amplifications, deletions, translocations, and dysregulation of noncoding and coding RNAs to provide a better understanding of the complex genetics and loss of regulation in cancer [15, 25, 29, 208, 209, 275]. For example, paired end sequencing showed that about half of structural rearrangements in breast cancer genomes were fusion transcripts resulting from the rearrangements of segmental tandem duplications involving multiple genes [277]. Similarly, other cancer types were found to be dominated by duplications, translocations, structural variations, and complex rearrangements called "chromothripsis" that involve chromosomal rearrangements as single events confined to genomic regions in one or a few chromosomes [278]. NGS also has been applied to circulating tumor cells isolated from the body fluids (blood, urine, sputum, saliva, and stools) [30, 274]. A genomic landscape and a catalogue of somatic mutations in cancer are provided on the Internet at COSMIC (Table 3, [275]). Thus, NGS potentially provides cancer patients with opportunities for personalized diagnosis and optimized therapeutic treatment [279, 280].

The integration of NGS data obtained from whole genome, exome, transcriptome, and methylome to build up individual genomic profiles is a growing reality in human health care. Recently, Chen et al. [269] developed "integrated personal omic profiling" in an individual by sequencing their genome at high accuracy and profiling their transcriptome, metabolome, and proteome over a 14-month period. In the study, they tracked the emergence of type 2 diabetes and assessed the individual's genetic make-up and disease risks. Others have performed similar studies demonstrating that monitoring the longitudinal trends and changes within individuals is an important future protocol for the diagnosis, management, and treatment of disease [9, 81, 270]. The challenges for "person omics," however, remain formidable at many levels, not least the time, cost, and effort required to gather, process, and interpret the data [101]. The cost benefits of NGS for personomics have still to be assessed with many economic, securities, personal, familial, social, and ethical issues to be considered and resolved.

9. Futuromics

The first-generation sequencing technologies and the pioneering computing and bioinformatics tools produced the initial sequencing data and information within a framework of structural and functional genomics in readiness for the following NGS developments. NGS provides substantially cheaper, friendlier, and more flexible high-throughput sequencing options with a quantum leap towards the generation of much more data on genomics, transcriptomics, and methylomics that translate more productively into proteomics, metabolomics, and systeomics. This major progression towards a more comprehensive characterization of genomes, epigenomes, and transcriptomes of humans and other species provides even more data as a proxy to probe diverse molecular interactions in the era of "omics" in many fields of biology, industry, and health care. A few years ago, the McKinsey Global Institute produced a report predicting that NGS and genomics, including the sequencing of a million human genomes,

would become an economically and socially disruptive technology as well as an annual trillion dollar industry by 2025 [281]. The authors assessed that next-generation genomics would affect many high impact areas of molecular biology and bioindustry such as improving genetic engineering tools to custom build organisms, genetically engineer biofuels, modify crops to improve farming practices and food stocks, and develop drugs to treat cancers and other diseases. Although these technologies promise huge benefits, they also come with social, ethical, and regulatory risks in regard to privacy and security of personal genetic information, the dangerous effects of modified organisms on the environment, the spectre of bioterrorism, eugenics, and concerns about the ownership and commercialization of genomic information. The application of prenatal genome sequencing for genetic screening already points to the potential of producing genetically modified babies with desired traits. Much will need to be done to educate and inform regulators and society about the risks and benefits when formulating the regulatory policies about the advances and applications of these next-generation technologies.

Today, NGS is the science of biological information systems and "Big Data,", but many challenges still remain in regard to NGS data acquisition, storage, analysis, integration, and interpretation [282, 283]. Future advancements will undoubtedly rely on new technologies and large-scale collaborative efforts from multidisciplinary and international teams to continue generating comprehensive, high-throughput data production and analysis. The availability of economically friendlier bench-top sequencers and third-generation sequencing tools will allow smaller laboratories and individual scientists to participate in the genomics revolution and contribute new knowledge to the different fields of structural and functional genomics in the life sciences. The authors of the following chapters in this book present additional examples, more detailed information, and a broader view of the methods and many advances, applications, and challenges of NGS that were either missed or not covered adequately in this opening chapter, particularly in regard to the RNA sequencing and transcriptome methods and data that provide us with a better understanding of functional genomics in microorganisms, plants, animals, and humans. *Te volo, bonam lectionem*.

Author details

Jerzy K. Kulski[1,2*]

Address all correspondence to: kulski@me.com

1 Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Isehara, Japan

2 Centre for Forensic Science, The University of Western Australia, Nedlands, WA, Australia

References

[1] Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463–7. PMCID: PMC431765

[2] Mardis ER: Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387–402. DOI: 10.1146/annurev.genom.9.081307.164359

[3] Mardis ER: A decade's perspective on DNA sequencing technology. Nature. 2011;470:198–203. DOI: 10.1038/nature09796

[4] Metzker ML: Sequencing technologies — The next generation. Nat Rev Genet. 2010;11:31–46. DOI: 10.1038/nrg2626

[5] Thompson JF, Milos PM: The properties and applications of single-molecule DNA sequencing. Genome Biol. 2011;12:217. DOI: 10.1186/gb-2011-12-2-217

[6] Margulies M, Egholm M, Altman WE, et al: Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80. PMID: 16056220

[7] Levy S, Sutton G, Ng PC, et al: The diploid genome sequence of an individual human. PLoS Biol. 2007;5:e254. PMID: 17803354

[8] Wheeler DA, Srinivasan M, Egholm M, et al: The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452:872–6. DOI: 10.1038/nature06884

[9] Gudbjartsson DF, Helgason H, Gudjonsson SA, et al: Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47:435–44. DOI: 10.1038/ng.3247

[10] The 1000 Genomes Project Consortium: An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65. DOI: 10.1038/nature11632

[11] Lam HY, Clark MJ, Chen R, et al: Performance comparison of whole-genome sequencing platforms. Nat Biotechnol. 2012;30:78–83. DOI: 10.1038/nbt.2065

[12] Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63. DOI: 10.1038/nrg2484

[13] Rabbini B, Tekin M, Mahdieh N: The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59:5–15. DOI: 10.1038/jhg.2013.114

[14] Leo VC, Morgan NV, Bern D, et al: Use of next-generation sequencing and candidate gene analysis to identify underlying defects in patients with inherited platelet function disorders. J Thromb Haemost. 2015;13:643–50. DOI: 10.1111/jth.12836

[15] Mardis ER, Wilson RK: Cancer genome sequencing: A review. Hum Mol Genet. 2009;18(R2):R163–8. DOI: 10.1093/hmg/ddp396

[16] Kulski JK, Suzuki S, Ozaki Y, Mitsunaga S, Inoko H, Shiina T: Phase HLA genotyping by next generation sequencing — A comparison between two massively parallel sequencing bench-top systems, the Roche GS Junior and Ion Torrent PGM. In: Xi Y, editor. HLA and Associated Important Diseases. Croatia: Intech; 2014. p. 141–81.

[17] Pelizzola M, Ecker JR: The DNA methylome. FEBS Lett. 2011;585:1994–2000. DOI: 10.1016/j.febslet.2010.10.061

[18] Ozsolak F, Milos PM: RNA sequencing: Advances, challenges and opportunities. Nat Rev Genet. 2011;12:87–98. DOI: 10.1038/nrg2934

[19] Wang K, Kim C, Bradfield J, et al: Whole-genome DNA/RNA sequencing identifies truncating mutations in RBCK1 in a novel Mendelian disease with neuromuscular and cardiac involvement. Genome Med. 2013;5:67. DOI: 10.1186/gm471

[20] Li S, Tighe SW, Nicolet CM, et al: Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF next generation sequencing study. Nat Biotechnol. 2014;32:915–25. DOI: 10.1038/nbt.2972

[21] Kulski JK, Kenworthy W, Bellgard M, et al: Gene expression profiling of Japanese psoriatic skin reveals an increased activity in molecular stress and immune response signals. J Mol Med (Berl). 2005;83:964–75. PMID: 16283139

[22] Meynert AM, Ansari M, FitzPatrick D, Taylor MS: Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinform. 2014;15:247. DOI: 10.1186/1471-2105-15-247

[23] Chang G, Gao S, Hou X, et al: High-throughput sequencing reveals the disruption of methylation of imprinted gene in induced pluripotent stem cells. Cell Res. 2014;24:293–306. DOI: 10.1038/cr.2013.173

[24] Ekram MB, Kim J: High-Throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS): Genome-wide DNA methylation analysis of IAP LTR retrotransposon. PLoS One. 2014;9:e101683. DOI: 10.1371/journal.pone.0101683

[25] Bakshi A, Ekram MB, Kim J: Locus-specific DNA methylation analysis of retrotransposons in ES, somatic and cancer cells using high-throughput targeted repeat element bisulphite sequencing. Genomics Data. 2015;3:87–9. PMID: 25554740

[26] Corley MJ, Zhang W, Zheng X, Lum-Jones A, Maunakea AK: Semiconductor-based sequencing of genome-wide DNA methylation states. Epigenetics. 2015;10:2:153–66. DOI: 10.1080/15592294.2014.1003747

[27] Farlik M, Sheffield NC, Nuzzo A, et al: Single-cell DNA methylome sequencing and bioinformatics inference of epi-genomic cell-state dynamics. Cell Rep. 2015;10:1386–97. DOI: 10.1016/j.celrep.2015.02.001

[28] Hirst M: Epigenomics: Sequencing the methylome. Methods Mol Biol. 2013;973:39–54. DOI: 10.1007/978-1-62703-281-0_3

[29] Li Y, Zhang Y, Li S, et al: Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer. Sci Rep. 2015;5:8790. DOI: 10.1038/srep08790

[30] Soon WW, Hariharan M, Snyder MP: High-throughput sequencing for biology and medicine. Mol Syst Biol. 2013;9:640. DOI: 10.1038/msb.2012.61.

[31] Watson JD, Crick FH: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171:737–8. PMID: 13054692

[32] Holley RW, Apgar J, Everett GA, et al: Structure of a ribonucleic acid. Science. 1965;147:1462–5. PMID: 5898068

[33] RajBhandary UL, Kohrer C: Early days of tRNA research: Discovery, function, purification and sequence analysis. J BioSci. 2006;31:439–51. PMID: 17206064

[34] Barba M, Czosnek H, Hadidi A: Historic perspective, development and applications of next-generation sequencing in plant virology. Viruses. 2014;6:106–36. DOI: 10.3390/v6010106

[35] Franca LTC, Carrilho E, Kist TBL: A review of DNA sequencing techniques. Q Rev Biophys. 2002:35:169–200. DOI: 10.1017/S0033583502003797

[36] Guzvic M: The history of DNA sequencing. J Med Biochem. 2013;32:301–12. DOI: 10.2478/jomb-2014-0004

[37] Maxam AM, Gilbert W: A new method for sequencing DNA. Proc Natl Acad Sci USA. 1977;74(2):560–4. PMID: 265521

[38] Smith LM, Sanders JZ, Kaiser RJ, et al: Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321:674–9. PMID: 3713851

[39] Saiki RK, Scharf S, Faloona F, et al: Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4. PMID: 2999980

[40] Temin HM, Mizutani S: RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature. 1970;226:1211–3. DOI: 10.1038/2261211a0

[41] Baltimore D: Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature. 1970;226:1209–11. PMID: 4316300

[42] Adams MD, Kelley JM, Gocayne JD, et al: Complementary DNA sequencing: Expressed sequence tags and human genome project. Science. 1991;252:1651–6. DOI: 10.1126/science.2047873

[43] International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. PMID: 11237011

[44] Venter JC, Adams MD, Myers EW, et al: The sequence of the human genome. Science. 2001;291:1304–51. PMID: 11181995

[45] Fleischmann RD, Adams MD, White O, et al: Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. Science. 1995;269:496–512. PMID: 7542800

[46] Fraser CM, Gocayne JD, White O, et al: The minimal gene complement of *Mycoplasma genitalium*. Science. 1995;270:397–404. PMID: 7569993

[47] Sutton GG, White O, Adams MD, Kerlavage AR: TIGR assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci Technol. 1995;1:9–19.

[48] Stein L: Genome annotation. From sequence to biology. Nat Rev Genet. 2001;2:493–503. PMID: 11433356

[49] Peltonen L, McKusick VA: Dissecting human disease in the postgenomic era. Science. 2001;291:1224–9. PMID: 11233446

[50] Kiechle FL, Zhang X: The postgenomic era. Implications for the clinical laboratory. Arch Pathol Lab Med. 2002;126:255–62. PMID: 11860296

[51] Zhu T: Global analysis of gene expression using GeneChip microarrays. Curr Opin Plant Biol. 2003;6:418–25. PMID: 12972041

[52] Lenoir T, Giannella E: Case study. The emergence and diffusion of DNA microarray technology. J Biomed Discov Collab. 2006;1:11. DOI: 10.1186/1747-5333-1-11

[53] Wetterstrand K: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [Internet]. Available from: https://www.genome.gov/sequencingcosts [Accessed: 2015-07-13].

[54] Liu L, Li Y, Li S, et al: Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364. DOI: 10.1155/2012/251364

[55] Head SR, Komori HK, LaMere SA, et al: Library construction for next-generation sequencing: Overviews and challenges. BioTech. 2014;56:61–77. DOI: 10.2144/000114133

[56] Hart C, Lipson D, Ozsolak F, et al: Single-molecule sequencing: sequence methods to enable accurate quantitation. Methods Enzymol. 2010;472:407e430. DOI: 10.1016/S0076-6879(10)72002-4

[57] Nextera XT DNA sample preparation guide [Internet]. Available from: http://www.liai.org/files/nextera_xt_sample_preparation_guide_15031942_c.pdf [Accessed: 2015-07-01].

[58] Illumina cBot [Internet]. Available from: http://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf (Accessed: 2015-07-01].

[59] Ion ChefTM or the Ion OneTouchTm 2 [Internet]. Available from: http://www.lifetechnologies.com/au/en/home/brands/ion-torrent.html [Accessed: 2015-07-01].

[60] Rothberg JM, Hinz W, Rearick TM, et al: An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475:348–52. DOI: 10.1038/nature10242

[61] Bentley DR, Balasubramanian S, Swerdlow HP, et al: Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9. DOI: 10.1038/nature07517

[62] Ma Z, Lee RW, Li B, et al: Isothermal amplification method for next-generation sequencing. Proc Natl Acad Sci USA. 2013;110:14320–3. DOI: 10.1073/pnas.1311334110

[63] Aird D, Ross MG, Chen WS, et al: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12:R18. DOI: 10.1186/gb-2011-12-2-r18

[64] Kozarewa, I, Kozarewa I, Ning Z, et al: Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009;6:291–5. DOI: 10.1038/nmeth.1311

[65] Bio-IT World Staff. Six Years After Acquisition, Roche Quietly Shutters 454 [Internet]. Available from: http://www.bio-itworld.com/2013/10/16/six-years-after-acquisition-roche-quietly-shutters-454.html [Accessed: 2015-06-16].

[66] Shendure J, Ji H: Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–45. DOI: 10.1038/nbt1486

[67] Balasubramanian S: Solexa sequencing: Decoding genomes on a population scale. Clin Chem. 2015;61:21–4. DOI: 10.1373/clinchem.2014.221747

[68] Illumina Sequencer Comparison Table [Internet]. Available from: http://www.illumina.com/systems/sequencing.html [Accessed: 2015-06-16].

[69] McCoy RC, Taylor RW, Blauwkamp TA, et al: Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. PLoS One. 2014;9(9):e106689. DOI: 10.1371/journal.pone.0106689

[70] Drmanac R, Sparks AB, Callow MJ, et al: Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327:78–81. DOI: 10.1126/science.1181498

[71] Retrovolocity [Internet]. Available from: http://www.completegenomics.com/revolocity/ [Accessed: 2015-06-26].

[72] Wang Y, Wen Z, Shen J, et al: Comparison of the performance of Ion Torrent chips in noninvasive prenatal trisomy detection. J Hum Genet. 2014;59:393–6. DOI: 10.1038/jhg.2014.40

[73] Diekstra A, Bosgoed E, Rikken A, et al: Translating Sanger-based routine DNA diagnostics into generic massive parallel ion semiconductor sequencing. Clin Chem. 2015;61:154–62. DOI: 10.1373/clinchem.2014.225250

[74] Schadt EE, Turner S, Kasarskis A: A window into third-generation sequencing. Hum Mol Gene. 2010;19:R227–40. DOI: 10.1093/hmg/ddq416

[75] Eid J, Fehr A, Gray J, et al: Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133–8. DOI: 10.1126/science.1162986

[76] Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW: A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 2010;38(15):e159. DOI: 10.1093/nar/gkq543

[77] Koren S, Harhay GP, Smith TP, et al: Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 2013;14:R101.

[78] Thompson JF, Steinmann KE: Single molecule sequencing with a HeliScope Genetic Analysis System. Curr Protoc Mol Biol. 2010;Chapter 7:Unit7.10. DOI: 10.1002/0471142727.mb0710s92

[79] Fuller CW, Middendorf LR, Benner SA, et al: The challenges of sequencing by synthesis. Nat Biotechnol. 2009;27:1013–23. DOI: 10.1038/nbt.1585

[80] Pushkarev D, Neff NF, Quake SR: Single-molecule sequencing of an individual human genome. Nat Biotechnol. 2009;27:847–52. PubMed: 19668243

[81] Ashley EA, Butte AJ, Wheeler MT, et al: Clinical assessment incorporating a personal genome. Lancet. 2010;375:1525–35. DOI: 10.1016/S0140-6736(10)60452-7

[82] Hickman SE, Kingery ND, Ohsumi T, et al: The microglial sensome revealed by direct RNA sequencing. Nat Neurosci. 2013;16:1896–905. DOI: 10.1038/nn.3554

[83] Bayley H: Nanopore sequencing: From imagination to reality. Clin Chem. 2015;61:25–31. DOI: 10.1373/clinchem.2014.223016

[84] Wang Y, Yang Q, Wang Z: The evolution of nanopore sequencing. Front Genet. 2005;5:1–20. DOI: 10.3389/fgene.2014.00449

[85] Bayley H, Cremer PS: Stochastic sensors inspired by biology. Nature. 2001;413:226–30. PMID: 11557992

[86] Stoddart D, Heron A, Mikhailova E, Maglia G, Bayley H: Single nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc Natl Acad Sci USA. 2009;106:7702–7. DOI: 10.1073/pnas.0901054106

[87] Laver T, Harrison J, O'Neill PA, et al: Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quant. 2015;3:1–8.

[88] Bell DC, Thomas WK, Murtagh KM, et al: DNA base identification by electron microscopy. Microsc Microanal. 2012;18:1049–53. DOI: 10.1017/S1431927612012615

[89] Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW: Shining a light on dark sequencing: Characterising errors in Ion Torrent PGM Data. PLoS Comput Biol. 2013;9:e1003031. DOI: 10.1371/journal.pcbi.1003031

[90] Gilles A, Meglecz E, Pech N, et al: Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011;12:245. DOI: 10.1186/1471-2164-12-245

[91] Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007;8:R143. PMID: 17659080

[92] Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32. DOI: 10.1038/nrg3642

[93] Loman NJ, Misra RV, Dallman TJ, et al: Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30:434–9. DOI: 10.1038/nbt.2198. Erratum in Nat Biotechnol. 2012;30:562.

[94] Quail M, Smith M, Coupland P, et al: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:341. DOI: 10.1186/1471-2164-13-341

[95] Rieber N, Zapatka M, Lasitschka B, et al: Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One. 2013;8:e66621. DOI: 10.1371/journal.pone.0066621

[96] Fuellgrabe MW, Herrmann D, Knecht H, et al: High-throughput, amplicon-based sequencing of the CREBBP gene as a tool to develop a universal platform-independent assay. PLoS One. 2015;10:e0129195. DOI: 10.1371/journal.pone.0129195

[97] Ozaki Y, Suzuki S, Kashiwase K, et al: Cost-efficient multiplex PCR for routine genotyping of up to nine classical HLA loci in a single analytical run of multiple samples by next generation sequencing. BMC Genomics. 2015;16:318. DOI: 10.1186/s12864-015-1514-4

[98] Horner DS, Pavesi G, Castrignano T, et al: Bioinformatics approaches for genomics and post genomics applications of next generation sequencing. Brief Bioinform. 2010;11:181–97. DOI: 10.1093/bib/bbp046

[99] El-Metwally S, Hamza T, Zakaria M, Helmy M: Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol. 2013;9:e1003345. DOI: 10.1007/s10142-015-0433-4

[100] Land M, Hauser L, Jun SR, et al: Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15:141–61. DOI: 10.1007/s10142-015-0433-4

[101] Hong HX, Zhang WQ, Shen J, et al: Critical role of bioinformatics in translating huge amounts of next-generation sequencing data into personalized medicine. Sci China Life Sci. 2013;56:110–8. DOI: 10.1007/s11427-013-4439-7

[102] Oliver GR, Hart SN, Klee EW: Bioinformatics for clinical next generation sequencing. Clin Chem. 2015;61:124–35. DOI: 10.1373/clinchem.2014.224360

[103] Yandell M, Ence D: A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13:329–42. DOI: 10.1038/nrg3174

[104] Schlotterer C, Tablerm R, Kofler R, Nolte V: Sequencing pools of individuals—Mining genome wide polymorphism data without big funding. Nat Rev Genet. 2014;15:749–63. DOI: 10.1038/nrg3803

[105] Shumway M, Cochrane G, Sugawara H: Archiving next generation sequencing data. Nucleic Acids Res. 2010;38:D870–1. DOI: 10.1093/nar/gkp1078

[106] Korneliussen TS, Albrechtsen A, Nielsen R: ANGSD: Analysis of next generation sequencing data. BMC Bioinform. 2014;15:356. DOI: 10.1186/s12859-014-0356-4

[107] Ruan J, Jiang L, Chong Z, et al: Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology. BMC Genomics. 2013;14:711. DOI: 10.1186/1471-2164-14-711

[108] Bradnam Fass JN, Alexandrov A, et al: Assemblathon 2: Evaluating *de novo* methods of genome assembly in three vertebrate species. Gigascience. 2013;2:10. DOI: 10.1186/2047-217X-2-10

[109] Medina I, Salavert F, Sanchez R, et al: Genome Maps, a new generation genome browser. Nucleic Acids Res. 2013;41:W41–6. DOI: 10.1093/nar/gkt530

[110] Cunningham F, Amode MR, Barrell D, et al: Ensembl 2015. Nucleic Acids Res. 2015;43:D662–9. DOI: 10.1093/nar/gku1010

[111] Fiers W, Contreras R, Duerinck F, et al: Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. 1976;260:500–7. PMID: 1264203

[112] Kuska B: Beer, Bethesda and biology: How "genomics" came into being. J Natl Cancer Inst. 1998;80:93. PMID: 9450566

[113] Hocquette JF, Cassar-Malek, Scalbert A, Guillou F: Contribution of genomics to the understanding of physiological functions. J Physiol Pharmacol. 2009;60(Suppl 3):5–16. PMID: 19996478

[114] Wylie KM, Weinstock GM, Storch GA: Virome genomics: A tool for defining the human virome. Curr Opin Microbiol. 2013;16(4):479–84. DOI: 10.1016/j.mib.2013.04.006

[115] Kwong JC, McCallum, Sintchenko V, Howden BP: Whole genome sequencing in clinical and public health microbiology. Pathol. 2015;47:199–210. DOI: 10.1097/PAT.0000000000000235

[116] Chun J, Rainey FA: Integrating genomics into taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol. 2014;64:316–24. DOI 10.1099/ijs.0.054171-0

[117] Ladner JT, Beitzel, Chain PSG: Standards for sequencing viral genomes in the era of high-throughput sequencing. mBio. 2014;5:e01360–14. DOI: 10.1128/mBio.01360-14

[118] Galagan JE, Henn MR, Ma L-J, Cuomo CA, Birren B: Genomics of the fungal kingdom: Insights into eukaryotic biology. Genome Res. 2005;15:1620–31. PMID: 16339359

[119] Stajich JE, Harris T, Brunk BP, et al: FungiDB: an integrated functional genomics database for fungi. Nucleic Acids Res. 2012;40(Database issue):D675–81. DOI: 10.1093/nar/gkr918

[120] Kim KM, Park J-H, Bhattacharya D, Yoon HS: Applications of next-generation sequencing to unraveling the evolutionary history of algae. Int J Syst Evol Microbiol. 2014;64:333–45. DOI: 10.1099/ijs.0.054221-0

[121] Pareek CSP, Smoczynski R, Tretyn A: Sequencing technologies and genome sequencing. J Appl Genet. 2011;52:413–35. DOI: 10.1007/s13353-011-0057-x

[122] Ekblom R, Wolf JBW: A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014;7:1026–42. DOI: 10.1111/eva.12178

[123] Gonzaga-Jauregui C, Lupski JR, Gibbs RA: Human genome sequencing in health and disease. Annu Rev Med. 2012;63:35–61. DOI: 10.1146/annurev-med-051010-162644

[124] Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ: The importance of phase information for human genomics. Nat Rev Genet. 2011;12:215–23. DOI: 10.1038/nrg2950

[125] Green RC, Berg JS, Grody WW, et al: ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15:565–74. DOI: 10.1038/gim.2013

[126] Taylor JC, Martin HC, Lise S, et al: Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47:717–26. DOI: 10.1038/ng.3304

[127] Handsaker RE, Van Doren V, Berman JR, et al: Large multiallelic copy number variations in humans. Nat Genet. 2015;47:296–303. DOI: 10.1038/ng.3200

[128] Ammar R, Paton TA, Torti D, Shlein A, Bader GD: Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Res. 2015;4:17. DOI: 10.12688/f1000research.6037.1

[129] Erlich HA: HLA typing using next generation sequencing: An overview. Hum Immunol. 2015;pii: S0198-8859(15)00093-2. DOI: 10.1016/j.humimm.2015.03.001

[130] Shiina T, Hosomichi K, Inoko H, Kulski JK: The HLA genomic loci map: Expression, interaction, diversity and disease. J Hum Genet. 2009;54:15–39. DOI: 10.1038/jhg.2008.5

[131] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8. DOI: 10.1038/nmeth.1226

[132] Nagalakshmi U, Wang Z, Waern K, et al: The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320:1344–9. DOI: 10.1126/science.1158441

[133] Kapranov P, St. Laurent G: Dark matter RNA: Existence, function, and controversy. Front Genet. 2012;3:article 60:1–9. DOI: 10.3389/fgene.2012.00060

[134] Birney E. Stamatoyannopoulos, JA, Dutta A, et al: Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. Nature. 2007;447:799–816. PMID: 17571346

[135] Clamp, M. Fry B, Kamal M, et al: Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA. 2007;104:19428–33. PMID: 18040051

[136] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18:1509–17. DOI: 10.1101/gr.079558.108

[137] Costa V, Angelini C, de Feis I, Ciccodicola A: Uncovering the complexity of transcriptomes with RNA-seq. J. Biomed. Biotechnol. 2010;2010:853916. DOI: 10.1155/2010/853916

[138] The RNAcentral Consortium: RNAcentral: an international database of ncRNA sequences. Nucleic Acids Res. 2015;43:D123–9. DOI: 10.1093/nar/gku991

[139] Huntzinger E, Izaurrralde E: Gene silencing by microRNAs: contributions of translational repression and miRNA decay. Nat Rev Genet. 2011;12:99–110. DOI: 10.1038/nrg2936

[140] Ross RJ, Weiner MM, Lin H: PIWI proteins and PIWI-interacting RNAs in the soma. Nature. 2014;505:353–9. DOI: 10.1038/nature12987

[141] Lindblad-Toh K, Garber M, Zuk O: A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478:476–82. DOI: 10.1038/nature10530

[142] Fujii YR: RNA genes: Retroelements and virally retroposable microRNAs in human embryonic stem cells. Open Virol J. 2010;4:63–75. DOI: 10.2174/1874357901004010063

[143] Kelley D, Rinn J: Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13:R107. DOI: 10.1186/gb-2012-13-11-r107

[144] Amaral PP, Dinger ME, Mattick JS: Non-coding RNAs in homeostasis, disease and stress responses: An evolutionary perspective. Brief Funct Genomics. 2013;12:254–78. DOI: 10.1093/bfgp/elt016

[145] Moolhuijzen P, Kulski JK, Dunn DS, et al: The transcript repeat element: The human Alu sequence as a component of gene networks influencing cancer. Funct Integr Genomics. 2010;10:307–19. DOI: 10.1007/s10142-010-0168-1

[146] Costa FF: Non-coding RNAs, epigenetics and complexity. Gene. 2008;410:9–17. DOI: 10.1016/j.gene.2007.12.008

[147] Derrien T, Johnson R, Bussotti G, et al: The GENCODEv7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution and expression. Genome Res. 2012;22:1775–89. DOI: 10.1101/gr.132159.111

[148] Yarmishyn AA, Kurochkin IV: Long noncoding RNAs: a potential novel class of cancer biomarkers. Front Genet. 2015;6:Article 145:1–10. DOI: 10.3389/fgene.2015.00145

[149] Iyer MK, Niknafs YS, Malik R: The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47:199–208. DOI: 10.1038/ng.3192

[150] Vance KW, Ponting CP: Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet. 2014;30:348–55. DOI: 10.1016/j.tig.2014.06.001

[151] Ward AJ, Cooper TA: The pathobiology of splicing. J Pathol. 2010;220:152–63. DOI: 10.1002/path.2649

[152] Poulos MG, Batra R, Charizanis K, Swanson MS: Developments in RNA splicing and disease. Cold Spring Harb Perspect Biol. 2011;3:a000778. DOI: 10.1101/cshperspect.a000778

[153] Melé M, Ferreira PG, Reverter F, et al: Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348:660–5. DOI: 10.1126/science.aaa0355

[154] Landt SG, Marinov GK, Kundaje A, et al: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012; 22:1813–31. DOI: 10.1101/gr.136184.111

[155] de Witt E, de Laat W: A decade of 3C technologies: Insights into nuclear organization. Genes Dev. 2012;26:11–24. DOI: 10.1101/gad.179804.111

[156] Berggard T, Linse S, James P: Methods for the detection and analysis of protein-protein interactions. Proteomics. 2007;7:2833–42. DOI: 10.1002/pmic.200700131

[157] Malm EK, Srivastava V, Sundqvist G, Bulone V: APP: An Automated Proteomics Pipeline for the analysis of mass spectrometry data based on multiple open access tools. BMC Bioinform. 2014;15:441. DOI: 10.1186/s12859-014-0441-8

[158] Chagoyen M, Pazos F: Tools for functional interpretation of metabolomics experiments. Brief Bioinform. 2013;14:737–44. DOI: 10.1093/bib/bbs055

[159] Pavlidid C, Patrinos GP, Katsila T: Nutrigenomics: A controversy. Appl Transl Genomics. 2015;4:50–3. DOI: 10.1016/j.atg.2015.02.003

[160] Winter G, Krömer JO: Fluxomics — Connecting "omics" analysis and phenotypes. Environ Microbiol. 2013;15:1901–16. DOI: 10.1111/1462-2920.12064

[161] Stanberry L, Mias GI, Haynes W, Higdon R, Snyder M, Kolker E: Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. Metabolites. 2013;3:741–60. DOI: 10.3390/metabo3030741

[162] Hernández-Prieto MA, Semeniuk TA, Futschik ME: Toward a systems-level understanding of gene regulatory, protein interaction, and metabolic networks in cyanobacteria. Front Genet. 2014;5:191. DOI: 10.3389/fgene.2014.0019

[163] Wang Q, Lu Q, Zhao H: A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet. 2015:6:149. DOI: 10.3389/fgene.2015.00149

[164] Editorial: Marshaling the variome. Nat Genet. 2015;47:849. DOI: 10.1038/ng.3377

[165] Groza T, Kohler S, Moldenhauer D, et al: The human phenotype ontology: Semantic unification of common and rare disease. Am J Hum Genet. 2015;97:111–24. DOI: 10.1016/j.ajhg.2015.05.020

[166] Paltoo DN, Rodriguez LL, Feolo M, et al: Data use under the NIH GWAS data sharing policy and future directions. Nat Genet. 2014;46:934–8. DOI: 10.1038/ng.3062

[167] Lappalainen I, Almeida-King J, Kumanduri V, et al: The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015;47:692–5. DOI: 10.1038/ng.3312

[168] Bashir Y, Singh SP, Konwar BK: Metagenomics: an application based perspective. Chin J Biol. 2014;ID146030:7 pages, http://dx.doi.org/10.1155/2014/146030

[169] Gilbert JA, Dupont CL: Microbial metagenomics: Beyond the genome. Annu Rev Mar Sci. 2011;3:347–71. PMID: 21329209

[170] Croucher NJ, Harris SR, Grad YH, Hanage WP: Bacterial genomes in epidemiology — Present and future. Phil Trans R Soc B. 2013;368:20120202. http://dx.doi.org/10.1098/rstb.2012.0202

[171] Grad YH, Lipsitch M: Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome Biol. 2014;15:538. http://genomebiology.com/2014/15/11/538

[172] Sunagawa S, Coelho LP, Chaffron S, et al: Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348:1261359. DOI: 10.1126/science.1261359

[173] de Vargas C, Audic S, Henry N, et al: Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348:1261605. DOI: 10.1126/science.1261605

[174] Brum JR, Ignacio-Espinoza JC, Roux S, et al: Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015;348:1261498. DOI: 10.1126/science.1261498

[175] Weinstock GM: Genomic approaches to studying the human microbiota. Nature. 2012;489:250–6. DOI: 10.1038/nature11553

[176] Morgan XC, Segata N, Huttenhower C: Biodiversity and functional genomics in the human microbiome. Trends Genet. 2013;29:51–8. DOI: 10.1016/j.tig.2012.09.005

[177] Ursell LK, Metcalf JL, Parfrey LW, Knight R: Defining the human microbiome. Nutr Rev. 2012;70 Suppl 1:S38–44. DOI: 10.1111/j.1753-4887.2012.00493.x

[178] Reddy TBK, Thomas A, Stamatis D, et al: The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2014;97:111–24. DOI: 10.1093/nar/gku950

[179] Myer F, Paarmann D, D'Souza M, et al: The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform. 2008;9:386. DOI: 10.1186/1471-2105-9-386

[180] Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR: Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc Natl Acad Sci USA. 2015;112:6449-54. http://dx.doi.org/10.1073/pnas.1421834112 (2015)

[181] Olsen GJ, Woese C: Archael genomics: An overview. Cell. 1997;89:991–4. PMID: 9215619

[182] Hagen JB: Five kingdoms, more or less: Robert Whittaker and the broad classification of organisms. BioScience. 2012;62:67–74. DOI: 10.1525/bio.2012.62.1.11

[183] Pawlowski J: The new micro-kingdoms of eukaryotes. BMC Biol. 2013;11:40. DOI: 10.1186/1741-7007-11-40

[184] Adl SM, Simpson AGB, Lane CE, et al: The revised classification of eukaryotes. J Eukaryot Microbiol. 2012;59:429–93. DOI: 10.1111/j.1550-7408.2012.00644.x

[185] Villarreal LP: How viruses shape the tree of life. Future Virol. 2006;1:587–95.

[186] Moreira D, Lopez-Garcia P: Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol. 2009;7:306–11. DOI: 10.1038/nrmicro2108

[187] Philippe N, Legendre M, Doutre G, et al: Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science. 2013;341:281–6. DOI: 10.1126/science.1239181

[188] Legendre M, Bartoli J, Shmakova L, et al: Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci USA. 2014;111:4274–9. DOI: 10.1073/pnas.1320670111

[189] Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ: Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. 2012;3:329. DOI: 10.3389/fmicb.2012.00329

[190] Shah V, Zakrzewski M, Wibberg D, Eikmeyer F, Schlüter A, Madamwar D: Taxonomic profiling and metagenome analysis of a microbial community from a habitat contaminated with industrial discharges. Microb Ecol. 2013;66:533-50. DOI: 10.1007/s00248-013-0244-x

[191] Soucy SM, Huang J, Gogarten JP: Horizontal gene transfer: Building the web of life. Nat Rev Genet. 2015;16:472–82. DOI: 10.1038/nrg3962

[192] Makarova KS, Wolf YI, Koonin EV: Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013;41:4360–77. DOI: 10.1093/nar/gkt157

[193] Haft DH, Selengut J, Mongodin EF, Nelson KN: A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol. 2005;1:e60. DOI: 10.1371/journal.pcbi.0010060

[194] Louwen R, Staals RHJ, Endtz HP, van Baarlen P, van der Oost J: The role of CRISPR-Cas systems in virulence of pathogenic bacteria. Microbiol Mol Biol Rev. 2014;78:74–88. DOI: 10.1128/Mmbr. 00039- 13

[195] Makarova KS, Haft DH, Barrangou R, et al: Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011;9:467–77. DOI: 10.1038/nrmicro2577

[196] Shalem O, Sanjana NE, Zhang F: High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015;16:299–310. DOI: 10.1038/nrg3899

[197] Liao, HK, Gu Y, Diaz A, et al: Use of the CRISPR-Cas9 system as an intracellular defense against HIV-1 infection in human cells. Nat Commun. 2015;6:6413. DOI: 10.1038/ncomms7413

[198] Chowdhury C, Sinha S, Chun S, Yeates TO, Bobik TA: Diverse bacterial microcompartment organelles. Microbiol Mol Biol Rev. 2014;78:438–68. DOI: 10.1128/mmbr.00009-14

[199] Jorda J, Lopez D, Wheatley NM, Yeates TO: Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria. Protein Sci. 2013;22:179–95. DOI: 10.1002/pro.2196

[200] Vasu K, Nagaraja V: Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev. 2013;77:53–72. DOI: 10.1128/MMBR.00044-12

[201] Zilberman D: The evolving functions of DNA methylation. Curr Opin Plant Biol. 2008;11:554–9. DOI: 10.1016/j.pbi.2008.07.004

[202] Guy L, Ettema TJG: The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol. 2011;19:580–7. DOI: 10.1016/j.tim.2011.09.002

[203] Spang A, Saw JH, Jørgensen SL, et al: Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015;521:173–9. DOI: 10.1038/nature14447

[204] Margulis L: Origin of Eukaryotic Cells. New Haven, CT: Yale University Press; 1970.

[205] Gray MW: Mitochondrial evolution. Cold Spring Harb Perspect Biol. 2012;4:a01140. DOI: 10.1101/cshperspect.a011403

[206] Kabnick KS, Peattie DA: *Gardia*: A missing link between prokaryotes and eukaryotes. Am Sci. 1991;79:34–43.

[207] Wilkins AS, Holliday R: The evolution of meiosis from mitosis. Genetics. 2009;181:3–12. DOI: 10.1534/genetics.108.099762

[208] Otto SP: The evolutionary consequences of polyploidy. Cell. 2007;13:452–62. PMID: 17981114

[209] Davoli T, de Lange T: The causes and consequences of polyploidy in normal development and cancer. Annu Rev Cell Dev Biol. 2011;27:585–610. DOI: 10.1146/annurev-cellbio-092910-154234

[210] Madlung A: Polyploidy and its effect on evolutionary success: old questions visited with new tools. Heredity. 2013;110:99–104. DOI: 10.1038/hdy.2012.79

[211] Hindle MM, Martin SF, Noordally ZB, et al: The reduced kinome of *Ostreococcus tauri*: core eukaryotic signally components in a tractable model species. BMC Genomics. 2014;15:640. DOI: 10.1186/1471-2164-15-640

[212] Pellicer J, Fay MF, Leitch IJ: The largest eukaryotic genome of them all? Botanical J Linnean Soc. 2010;164:10–5. DOI: 10.1111/j.1095-8339.2010.01072.x

[213] Katju V, Bergthorsson U: Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front Genet. 2013;4:273. DOI: 10.3389/fgene.2013.00273

[214] Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M: Two evolutionary distinct classes of paleopolyploidy. Mol Biol Evol. 2014;31:448–54. DOI: 10.1093/molbev/mst230

[215] Dehal P, Boore JL: Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 2005;3:e314. DOI: 10.1371/journal.pbio.0030314

[216] Ohno S. Evolution by Gene Duplication. New York: Springer-Verlag; 1970.

[217] Parisod C, Mhiri C, Lim KY, et al: Differential dynamics of transposable elements during long-term diploidization of *Nicotiana* Section *Repandae* (Solanaceae) allopolyploid genomes. PLoS One. 2012;7:e50352. DOI: 10.1371/journal.pone.0050352

[218] Renny-Byfield S, Kovarik A, Kelly LJ, et al: Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. Plant J. 2013;74:829–39. DOI: 10.1111/tpj.12168

[219] Wang X, Freeling M: The Brassica genome. Front Plant Sci. 2013;4:148. DOI: 10.3389/fpls.2013.00148

[220] Shikari AB, Parray GA, Sofi NR, Hussain A, Dar ZA, Iqbal AM: Group balanced block design for comparisons among oilseed *Brassicae*. Sci Res Essays. 2015;10:302–5. DOI: 10.3389/fpls.2013.00148

[221] Michael TP, Jackson S: The first 50 plant genomes. Plant Genome. 2013;6:1–7.

[222] Feldman M, Levy AA: Genome evolution due to allopolyploidization in wheat. Genetics. 2012;192:763–74. DOI: 10.1534/genetics.112.146316

[223] Chapman, Mascher M, Buluc A, et al: A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 2015;16:26. DOI: 10.1186/s13059-015-0582-8

[224] Ainouche ML, Fortune PM, Salmon A, et al: Hybridization, polyploidy and invasion: Lessons from *Spartina* (Poaceae). Biol Inv. 2008;11:1159–73. DOI: 10.1007/s10530-008-9383-2

[225] Brawand D, Wagner CE, Li YI, et al: The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513:375–81. DOI: 10.1038/nature13726

[226] Seehausen O, Butlin RK, Keller I, et al: Genomics and the origin of species. Nat Rev Genet. 2014;15:176–92. DOI: 10.1038/nrg3644

[227] Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO: Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci USA. 2013;110:E1594–603. DOI: 10.1073/pnas.1211371110

[228] Feschotte C: The contribution of transposable elements to the evolution of regulatory networks. Nat Rev Genet. 2008;9:397–405. DOI: 10.1038/nrg2337

[229] Bejerano G, Pheasant M, Makunin I, et al: Ultraconserved elements in the human genome. Science. 2004;304:1321–5. DOI: 10.1126/science.1098119

[230] Miller W, Rosenbloom K, Hardison RC, et al: 28-way vertebrate alignment and conservation track in the UCSC genome browser. Genome Res. 2007;17:1797–808. PMID: 17984227

[231] Dimitrieva S, Bucher P: UCNEbase—A database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Res. 2013;41(Database issue):D101-9. DOI: 10.1093/nar/gks1092

[232] Luo H, Lin Y, Gao F, Zhang C-T, Zhang R: DEG 10, an update of the Database of Essential Genes that includes both protein-coding genes and non-coding genomic elements. Nucleic Acids Res. 2014;42:D574–80. DOI: 10.1093/nar/gkt1131

[233] Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T: The human phylome. Genome Biol. 2007;8:R109. PMID: 17567924

[234] Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T: PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014;42(Database issue):D897–902. DOI: 10.1093/nar/gkt1177

[235] Matsuzaka Y, Okamoto K, Mabuchi T, et al: Identification, expression analysis and polymorphism of a novel RLTPR gene encoding a RGD motif, tropomodulin domain and proline/leucine-rich regions. Gene. 2004;343:291–304. PMID: 15588584

[236] Liang Y, Cucchetti M, Roncagalli R, et al: The lymphoid lineage-specific actin-uncapping protein Rltpr is essential for costimulation via CD28 and the development of regulatory T cells. Nat Immunol. 2013;14:858–66. DOI: 10.1038/ni.2634

[237] Willems L, Gillet NA: APOBEC3 interference during replication of viral genomes. Viruses. 2015;7:2999–3018. DOI: 10.3390/v7062757

[238] Richard GF, Kerrest A, Dujon B: Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008;72:686–727. DOI: 10.1128/MMBR.00011-08

[239] McClintock B: The significance of responses of the genome to challenge. Science. 1984;226:792–801. PMID: 15739260

[240] Jurka J, Klonowski P, Dagman V, Pelton P: CENSOR—A program for identification and elimination of repetitive elements from DNA sequences. Comput Chem. 1996;20:119–21. PMID: 8867843

[241] Kohany O, Gentles AJ, Hankus L, Jurka J: Annotation, submission and screening of repetitive elements in Repbase: Rep baseSubmitter and Censor. BMC Bioinform. 2006;7:474. DOI: 10.1186/1471-2105-7-474

[242] Thung DT, de Ligt J, Vissers LEM, et al: Mobster: Accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 2014:15:488. PMID: 25348035

[243] Girgis HZ: Red: An intelligent, rapid, accurate tool for detecting repeats *de-novo* on the genomic scale. BMC Bioinform. 2015;16:227. DOI 10.1186/s12859-015-0654-5

[244] Tempel S, Talla E. VisualTE: a graphical interface for transposable element analysis at the genomic scale. BMC Genomics. 2015;16:139. DOI 10.1186/s12864-015-1351-5

[245] Burns KH, Boeke JD: Human transposon tectonics. Cell. 2012;149:740–52. DOI: 10.1016/j.cell.2012.04.019

[246] Feschotte C, Pritham EJ: DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331–48. PMID: 18076328

[247] Frost LS, Leplae R, Summers AO, Toussaint A: Mobile genetic elements: The agents of open source evolution. Nat Rev Microbiol. 2005;3:722–32. PMID:16138100

[248] Giordano J, Ge Y, Gelfand Y, Abrusa´n G, Benson G, et al: Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol. 2007;3:e137. DOI: 10.1371/journal.pcbi.0030137

[249] Hoen DR, Bureau TE: Discovery of novel genes derived from transposable elements using integrative genomic analysis. Mol Biol Evol. 2015;32(6):1487–506. DOI: 10.1093/molbev/msv042

[250] Bao W, Kojima KK, Kohany O: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015;6:11. DOI: 10.1186/s13100-015-0041-9

[251] Menconi G, Battaglia G, Grossi R, Pisanti N, Marangoni R: Mobilomics in *Saccharomyces cerevisiae* strains. BMC Bioinform. 2013;14:102. DOI: 10.1186/1471-2105-14-102

[252] SanMiguel P, Tikhonov A, Jin YK, et al: Nested retrotransposons in the intergenic regions of the maize genome. Science. 1996;274:765–8. PMID: 12021852

[253] Koonin EV, Wolf YI: Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36:6688–719. DOI: 10.1093/nar/gkn668

[254] Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: Whole-genome analysis of photosynthetic prokaryotes. Science. 2002;298:1616–20. PMID: 12446909

[255] Rankin DJ, Rocha EPC, Brown SP: What traits are carried on mobile genetic elements, and why? Heredity. 2011;106:1–10. DOI: 10.1038/hdy.2010.24

[256] Boto L: Horizontal gene transfer in the acquisition of novel traits by metazoans. Proc R Soc B. 2014;281:20132450. http://dx.doi.org/10.1098/rspb.2013.2450

[257] Jjingo D, Conley AB, Wang J, et al: Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. Mobile DNA. 2014;5:14. DOI: 10.1186/1759-8753-5-14

[258] Wang J, Vicente-García C, Seruggia D, et al: MIR retrotransposon sequences provide insulators to the human genome. Proc Natl Acad Sci USA. 2015;112:E4428–37. DOI: 10.1073/pnas.1507253112

[259] Kulski JK, Gaudieri S, Inoko H, Dawkins RL: Comparison between two human endogenous retrovirus (HERV)-rich regions within the major histocompatibility complex. J Mol Evol. 1999;48:675–83. PMID: 10229571

[260] Kulski JK, Gaudieri S, Martin A, Dawkins RL: Coevolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. J Mol Evol. 1999;49:84–97. PMID: 15269276

[261] Kulski JK, Anzai T, Shiina T, Inoko H: Rhesus macaque class I duplicon structures, organization, and evolution within the alpha block of the major histocompatibility complex. Mol Biol Evol. 2004;21:2079–91. PMID:15269276

[262] Lander E: Initial impact of the sequencing of the human genome. Nature. 2011;470:187–97. DOI: 10.1038/nature09792

[263] Ball MP, Bobe JR, Chuo MF, et al: Harvard Personal Genome Project: Lessons from participatory public research. Genome Med. 2014;6:10. DOI: 10.1186/gm527

[264] The International HapMap Consortium: Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52-58. DOI: 10.1038/nature09298

[265] HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, et al: Mapping human genetic diversity in Asia. Science. 2009;326:1541-5. DOI: 10.1126/science.1177074

[266] Pagini L, Schiffels S, Gurdasani D, et al: Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015;96:986–91. DOI: 10.1016/j.ajhg.2015.04.019

[267] Colonna V, Ayub Q, Chen Y, et al: Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 2014;14:R88. DOI: 10.1186/gb-2014-15-6-r88

[268] Johnson JJ, Lewis KL, Ng D, et al: Individualized iterative phenotyping for genome-wide analysis of loss-of-function mutations. Am J Hum Genet. 2015;96;913–25. DOI: 10.1016/j.ajhg.2015.04.013

[269] Chen R, Mias GI, Li-Pook-Than J, et al: Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148:1293–307. DOI: 10.1016/j.cell.2012.02.009

[270] Frese KS, Katus HA, Meder B: Next generation sequencing. From understanding biology to personalized medicine. Biology. 2013;2:378–98. DOI: 10.3390/biology2010378

[271] Veeramah KR, Hammer MF: The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet. 2014;15:162. DOI: 10.1038/nrg3625

[272] Hofreiter M, Paijmans LA, Goodchild H, et al: The future of ancient DNA: Technical advances and conceptual shifts. Bioessays. 2014;37:284–93. DOI: 10.1002/bies.201400160

[273] Tryka KA, Hao L, Sturcke A, et al: The Database of Genotypes and Phenotypes (dbGaP) and PheGenI. The NCBI Handbook [Internet]. 2nd ed., 2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK154410/ [Accessed: 2015-07-25].

[274] Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155:27–38. DOI: 10.1016/j.cell.2013.09.006

[275] Forbes SA, Bindal N, Bamford S, et al: COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39:D945–50. DOI: 10.1093/nar/gkq929

[276] Boyle AP, Araya CL, Brdlik C, et al: Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512:453–6. DOI: 10.1038/nature13668

[277] Inaki K, Hillmer AM, Ukil L, et al: Transcriptional consequences of genomic structural aberrations in breast cancer. Genome Res. 2011;21:676–87. DOI: 10.1101/gr.113225.110

[278] Zhang C-Z, Spektor A, Comils, et al: Chromothripsis from DNA damage in micronuclei. Nature. 2015;522:179–84. DOI: 10.1038/nature14493

[279] Mardis ER: Genome sequencing and cancer. Curr Opin Genet Dev. 2012;22:1–6. http://dx.doi.org/10.1016/j.gde.2012.03.005

[280] Jones S, Anagnostou V, Lytle K, et al: Personalized genomic analyses for cancer mutation discovery and interpretation. Sci Transl Med. 2015;7:283ra53. DOI: 10.1126/scitranslmed.aaa7161

[281] Manyika J, Chui M, Bughin J, Dobbs R, Bisson P, Marrs A: Disruptive technologies: Advances that will transform life, business, and the global economy. McKinsey Global Institute, May 2013. Available from: http://www.mckinsey.com/insights/business_technology/disruptive_technologies [Accessed: 2015-08-10].

[282] Tang H, Zhao Z: Bioinformatics drives the applications of next-generation sequencing in translational biomedical research. Methods. 2015;79–80:1–2. DOI: 10.1016/j.ymeth.2015.04.035

[283] Stephens ZD, Lee SY, Faghri F, et al: Big Data: Astronomical or genomical? PLoS Biol. 2015;13:e1002195. DOI: 10.1371/journal.pbio.1002195