Review

CellPress

Ten years of next-generation sequencing technology

Erwin L. van Dijk[1], Hélène Auger[1], Yan Jaszczyszyn[2], and Claude Thermes[1]

[1] Centre de Génétique Moléculaire – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France
[2] Plateforme Intégrée IMAGIF – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France

Ten years ago next-generation sequencing (NGS) technologies appeared on the market. During the past decade, tremendous progress has been made in terms of speed, read length, and throughput, along with a sharp reduction in per-base cost. Together, these advances democratized NGS and paved the way for the development of a large number of novel NGS applications in basic science as well as in translational research areas such as clinical diagnostics, agrigenomics, and forensic science. Here we provide an overview of the evolution of NGS and discuss the most significant improvements in sequencing technologies and library preparation protocols. We also explore the current landscape of NGS applications and provide a perspective for future developments.

An overview of ten years of next-generation sequencing technology

In the 1970s, Sanger and colleagues [1] and Maxam and Gilbert [2] developed methods to sequence DNA by chain termination and fragmentation techniques, respectively. This transformed biology by providing the tools to decipher complete genes and, later, entire genomes. The technique developed by Sanger and colleagues, commonly referred to as Sanger sequencing, required less handling of toxic chemicals and radioisotopes than Maxam and Gilbert's method, and as a result it became the prevailing DNA sequencing method for the next 30 years. A growing demand for increased throughput led to laboratory automation and process parallelization, which eventually resulted in the establishment of factory-like outfits with hundreds of sequencing instruments. Thanks to these advances, the Sanger technique ultimately enabled the completion of the first human genome sequence in 2004 [3].

The Human Genome Project (see Glossary), however, required vast amounts of time and resources and it was clear that faster, higher throughput, and cheaper technologies were required. For this reason, in the same year (2004) the National Human Genome Research Institute (NHGRI) initiated a funding program with the goal of reducing the cost of human genome sequencing to US$1000 in ten years [4]. This stimulated the development and commercialization of next-generation sequencing (NGS) technologies, as opposed to the automated Sanger method, which is considered a first-generation technology. These new sequencing methods share three major improvements. First, instead of requiring bacterial cloning of DNA fragments they rely on the preparation of NGS libraries in a cell free system. Second, instead of hundreds,

Corresponding author: van Dijk, E.L. (vandijk@cgm.cnrs-gif.fr, e_dijk@yahoo.com).
Keywords: Next-generation sequencing (NGS); DNA-seq; RNA-seq; ChIP-seq; NGS library preparation; genomics.

0168-9525/
 http://dx.doi.org/10.1016/j.tig.2014.07.001

Glossary

Base-calling software: software to analyze the raw data produced by automated sequencers to predict the individual bases.
Cell lineage tree: a mathematical entity that describes the history of the cells in an organism or tissue, from conception until any particular moment in time. The root of the tree represents the mother cell, the leaves of the tree represent the extant cells, and branches in the tree capture every single cell division in the organism's or tissue's history. The cell lineage tree of only one organism, *Caenorhabditis elegans*, is known.
Epigenetic mark: a feature not directly encoded in the genetic code, including methylation of DNA and covalent modification of histone proteins. The latter may be tagged with methyl, acetyl, ubiquitin, phosphate, poly(ADP)ribose, and other biochemical groups. These modifications may influence chromatin structure and gene expression.
Human Genome Project: International scientific research project to determine the human genome sequence and map all of its genes. The project was initiated in 1990 and a first rough draft of the genome was published in 2001. The complete genome was published in 2004. About 20 500 genes were identified, but with ongoing analyses of the sequences this number may still change (just recently, for example, this number was decreased to 19 000). The Human Genome Project is the world's largest collaborative biological project to date and has cost about US$3 billion.
Metagenomics: the study of organisms in a microbial community based on analyzing the DNA within an environmental sample. Environmental metagenomics data are used for agricultural microbiome analysis, ecological remediation, and other biological investigations. Human microbiome analysis is the study of microbial communities found in and on the human body. The goal of human microbiome studies is to understand the role of microbes in health and disease.
Microarray: a collection of oligonucleotides attached to a solid surface (chip). Microarrays are used for transcriptome analysis, genotyping, and identification of protein–DNA interactions or epigenetic patterns on a genome-wide scale (ChIP-chip). The oligonucleotides on the chip are used to hybridize fluorophore-, silver-, or chemiluminescence-labeled (c)DNA or cRNA (also called antisense RNA) libraries under high-stringency conditions.
Multiplexed run: NGS run in which sequencing libraries are mixed. The unique index of each library allows the user to distinguish the libraries and to dissect the mixed sequence reads into separate files with the reads belonging to each individual library in a process called demultiplexing.
Nuclease footprinting: a molecular biology technique that detects DNA–protein interactions by exploiting the fact that a protein bound to DNA will often protect that DNA from enzymatic cleavage by deoxyribonuclease (DNase) enzymes. This makes it possible to locate a protein-binding site on a particular DNA molecule.
Single nucleotide polymorphism (SNP): a single variable (i.e., polymorphic) nucleotide in the DNA sequence. SNPs are usually bi-allelic, but multi-allelic SNPs do exist. They are the most common form of variation in the genome and are used extensively to study genetic differences between individuals and populations.
Whole-exome sequencing (WES): method to sequence only the RNA coding regions of a genome, the exome (all exons in all genes). WES involves the capture of fragmented genomic DNA by oligonucleotide probes that collectively cover all exonic regions.

thousands-to-many-millions of sequencing reactions are produced in parallel. Third, the sequencing output is directly detected without the need for electrophoresis; base interrogation is performed cyclically and in parallel. The enormous numbers of reads generated by NGS enabled the sequencing of entire genomes at an unprecedented speed. However, a drawback of NGS technologies was their relatively short reads. This made genome assembly more difficult and required the development of novel alignment algorithms (see below). The first NGS technology to be released in 2005 was the pyrosequencing method by 454 Life Sciences (now Roche) [5]. The 454 Genome Sequencer generated about 200 000 reads (~20 Mb) of 110 base-pairs (bp). One year later, the Solexa/Illumina sequencing platform was commercialized (Illumina acquired Solexa in 2007). The third technology to be released was Sequencing by Oligo Ligation Detection (SOLiD) by Applied Biosystems (now Life Technologies) in 2007 [6]. The Illumina and SOLiD sequencers generated much larger numbers of reads than 454 (30 and 100 million reads, respectively) but the reads produced were only 35 bp long. In 2010, Ion Torrent (now Life Technologies) released the Personal Genome Machine (PGM). This system was developed by Jonathan Rothberg, the founder of 454, and resembles the 454 system. An important difference is that the PGM uses semiconductor technology and does not rely on the optical detection of incorporated nucleotides using fluorescence and camera scanning. This resulted in higher speed, lower cost, and smaller instrument size. The first PGM generated up to 270 Mb of sequence with up to 100 nt reads; slightly shorter than those produced by 454. For a detailed description of these major NGS technologies, see [7,8]; a quick overview is presented in Box 1.

Other NGS methods have been developed, such as Qiagen-intelligent bio-systems sequencing-by-synthesis [9], Polony sequencing [10], and a single molecule detection system (Helicos BioSciences) [11]. In this latter system, the template DNA is not amplified before sequencing, which places this method at the interface between NGS and the so-called third-generation sequencing technologies. Third-generation methods also allow the detection of single molecules but as an additional common feature sequencing occurs in real time [12]. The leader in this field is currently Pacific Biosciences (PacBio). Their first instrument, the PacBio RS, appeared in 2010 and generated several thousands of up-to-several-kilobase-long reads [13]. The long reads made this technology ideal for the completion of *de novo* genome assemblies. PacBio is based on the detection of natural DNA synthesis by a single DNA polymerase. Incorporation of phosphate-labeled nucleotides leads to base-specific fluorescence, which is detected in real time. Sequencing runs therefore last minutes or hours rather than days (http://www.pacificbiosciences.com/products/). Here, we focus on the five platforms that have dominated the NGS market over the past decade: 454, Illumina, SOLiD, Ion Torrent, and PacBio.

Box 1. NGS technologies

Sample preparation for 454, Ion Torrent, and SOLiD technology
NGS libraries (Box 3) are first captured on beads (one fragment per bead). A water-in-oil emulsion containing PCR reagents and one bead per droplet is created to amplify each fragment individually. Subsequently, DNA is denatured and the beads, containing one amplified DNA fragment each, are distributed into the wells of a fiberoptic slide (454 and Ion Torrent: one bead per well) or on a glass slide (SOLiD).

Sample preparation for Illumina/Solexa technology
Libraries are denatured and bound at one end to a solid surface coated with adapter oligonucleotides. The free end of each fragment 'bends over' and hybridizes to a complementary adapter on the surface, which initiates complementary strand synthesis. Multiple cycles of this solid-phase amplification followed by denaturation create clusters of ~1000 copies of single-stranded DNA molecules.

454: pyrosequencing
The wells are loaded with sequencing enzymes and primer, then exposed to a flow of one unlabeled nucleotide at a time, allowing synthesis of the complementary DNA strand. When a nucleotide is incorporated, pyrophosphate is released leading to light emission, which is monitored in real time.

Ion Torrent: semiconductor sequencing
Very similar to 454 sequencing except that, instead of pyrophosphate, proton release during nucleotide incorporation is detected using ion sensors; no imaging technology is required.

Illumina/Solexa: sequencing with reversible terminators
Synthesis reagents consist of primers, DNA polymerase, and four differently labelled, reversible terminator nucleotides. After incorporation of a nucleotide, which is identified by its color, the 3′ terminator on the base and the fluorophore are removed, and the cycle is repeated.

SOLiD: sequencing by ligation
A sequencing primer is hybridized to the adapter and its 5′ end is available for ligation to an oligonucleotide hybridizing to the adjacent sequence. A mixture of octamers compete for ligation to the primer (bases 4 and 5 in these oligos are encoded by one of four color labels). After color detection, the ligated octamer is cleaved between position 5 and 6, which removes the label, and the cycle is repeated. In the first round, the process determines possible identities of bases in positions 4, 5, 9, 10, 14, 15, etc. The process is repeated, offset by one base using a shorter sequencing primer, to determine positions 3, 4, 8, 9, 13, 14, etc., until the first base in the sequencing primer (position 0) is reached.

The revolution

The advent of NGS immediately revolutionized genomics research by bringing the sequencing of entire genomes within reach of many small laboratories. In addition, gene expression studies frequently changed from using microarrays to NGS-based methods, which enabled the identification and quantification of transcripts without prior knowledge of a particular gene and provided information regarding alternative splicing and sequence variation [14]. For genome-wide mapping of protein–DNA interactions and epigenetic marks, chromatin immunoprecipitation followed by sequencing (ChIP-seq) was another early application of NGS [15]. Microarrays had been used before, but ChIP-seq provided substantially improved data with higher resolution and a larger dynamic range. For these quantitative applications, Illumina and SOLiD sequencing were more suitable than 454, owing to their higher throughput. For this reason, transcriptome profiling and ChIP-seq studies have mostly used Illumina or SOLiD sequencing [14,15]. By contrast, the reads generated by these technologies were initially too short for *de novo* genome assemblies. Thus, for this type of application, 454 was the preferred technology and enabled exciting discoveries such as the first million bp of a Neandertal genome [16]. Another important application of 454 was metagenomics, for example uncovering the potential cause of the disappearance of the honeybee [17].

Through the upgrading of sequencing machines and improvements in base-calling software and in sequencing chemistries, Illumina technology can now generate reads of several hundreds of bp long (Figure 1A). Thus, although 454 still produces longer reads, *de novo* genome assembly and metagenomics can now also be performed with Illumina sequencing. Exceptionally long reads are produced by the new PacBio RS II with maximum read lengths of over 20 kb, making this technology an ideal tool to finish genome assemblies [18].

A particularly impressive increase in throughput has been achieved by Illumina, which currently offers the highest throughput per run and the lowest per-base cost [8] (Figure 1B). Recently, Illumina released the HiSeq X Ten, a set of ten HiSeq X sequencing machines, with the staggering capacity to generate up to 1.8 Tb of sequence per run (http://www.illumina.com). With this novel system Illumina claims to have broken the barrier of the US$1000 genome corresponding to the original goal of the NHGRI funding program. This would mean a 10 000-fold reduction in price relative to the cost of a human genome in 2004 (http://www.genome.gov/sequencingcosts/) (Figure 1C). It should be noted, however, that the US$1000 cost requires that all HiSeq X machines run at full capacity, thus delivering about 18 000 genomes a year. This fact, together with a total system cost of at least US$10 million, implies that the HiSeq X Ten system will only be available to large institutes performing population-scale genome sequencing.

How was this tremendous increase in throughput achieved? The HiSeq X contains four main improvements: (i) unlike the classical flow cells where DNA templates are randomly distributed, the HiSeq X uses patterned flow cells that contain billions of nanowells at fixed locations, enabling extremely high cluster density; (ii) a new clustering chemistry has been developed to achieve a high occupancy and monoclonality within each well; (iii) a faster camera; and (iv) new polymerase enzymes to perform faster sequencing reactions lead to shorter run times (http://res.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf).

The general technological improvements of NGS led to their widespread use, and there was a growing interest from the clinic to use NGS as a diagnostic tool. These applications required cheaper, faster, and easier-to-use sequencers and, to meet this demand, Roche and Illumina launched compact bench-top sequencers. In late 2009, Roche launched the GS Junior, which produces 35 Mb of 400 nt reads in 10 h. In 2011, Illumina released the MiSeq, which initially produced 1.5 Gb of 150 nt reads in 27 h (new reagents can generate 25 million 300 nt reads; i.e., 15 Gb), but the fastest bench-top NGS platform remains Ion Torrent's PGM (Figure 1D). The PacBio system is equally fast but is much more expensive and therefore less suitable for small laboratories and the clinical setting. A summary of the advantages and drawbacks of the different NGS systems is presented in Box 2.

Improvements in NGS sample preparation methods and data analysis algorithms

Sample preparation methods have rapidly evolved along with the sequencing technologies. In a typical NGS library, DNA or RNA molecules are fused with adapters that contain the necessary elements for immobilization on a solid surface and sequencing (Box 3). Major problems in NGS library preparation are the introduction of quantitative biases and the loss of material. However, creative solutions have been found to combat these issues. Because PCR is a major source of bias, systematic comparisons of PCR polymerases have been performed and have identified enzymes that introduce less noise than the traditional ones [19]. RNA sequencing (RNA-seq) protocols are particularly bias-prone and many other steps besides PCR such as random-primed cDNA synthesis or adaptor ligation can introduce error [20]. However, a comparison of strand-specific RNA-seq protocols revealed that a method incorporating the specific elimination of second strands of cDNA (dUTP method) performs relatively well [21]. Significant reduction of sample loss has been achieved through Nextera technology, which combines DNA fragmentation, end-polishing, and adaptor-ligation into a single reaction [22]. In addition, gel and column purification steps have been replaced by magnetic beads, further reducing loss of material [23]. These improvements translate into reductions in the amount of input material required; where initially micrograms were needed, today often nanogram quantities are sufficient. Further, advances in whole-genome amplification (WGA) methods and miniaturization of reaction volumes through microfluidics technology even enable the sequencing of genomic DNA or RNA extracted from single cells [24,25]. WGA remains technically challenging, however, and for high coverage sequencing of single cell genomes further technical improvements are needed. In addition to general improvements in sample preparation, many novel protocols for specific applications have been developed, which are discussed below.

It should be noted that the enormous increase in throughput of NGS has also been due in part to significant advances in data handling. The development of NGS has made huge demands on bioinformatics tools for data analysis and management. In addition, the relatively short reads compared with Sanger sequences required novel alignment algorithms. A large number of new algorithms specifically designed to manage short reads have been developed [26], as well as new algorithms for *de novo* sequence assembly [27], single nucleotide polymorphism (SNP) detection [28], ChIP-seq analysis [29], and RNA-seq analysis [30]. Algorithms have also been developed to correct for biases introduced during library preparation [31]. The combination of experimental and computational improvements has greatly increased the utility of NGS technologies, and these areas are likely to continue to maximize the amount of information that it is possible to glean from NGS platforms.

A wide diversity of novel NGS applications

Genomic DNA sequencing

Advances in throughput and cost reduction have made WGS at the population scale increasingly feasible. Since the first large-scale human genetic variation study, the 1000 Genomes Project [32], ever-larger projects have been launched, involving the sequencing of thousands [33] or even millions of genomes (http://www.genomics.cn/en/index). These projects are revolutionizing our understanding

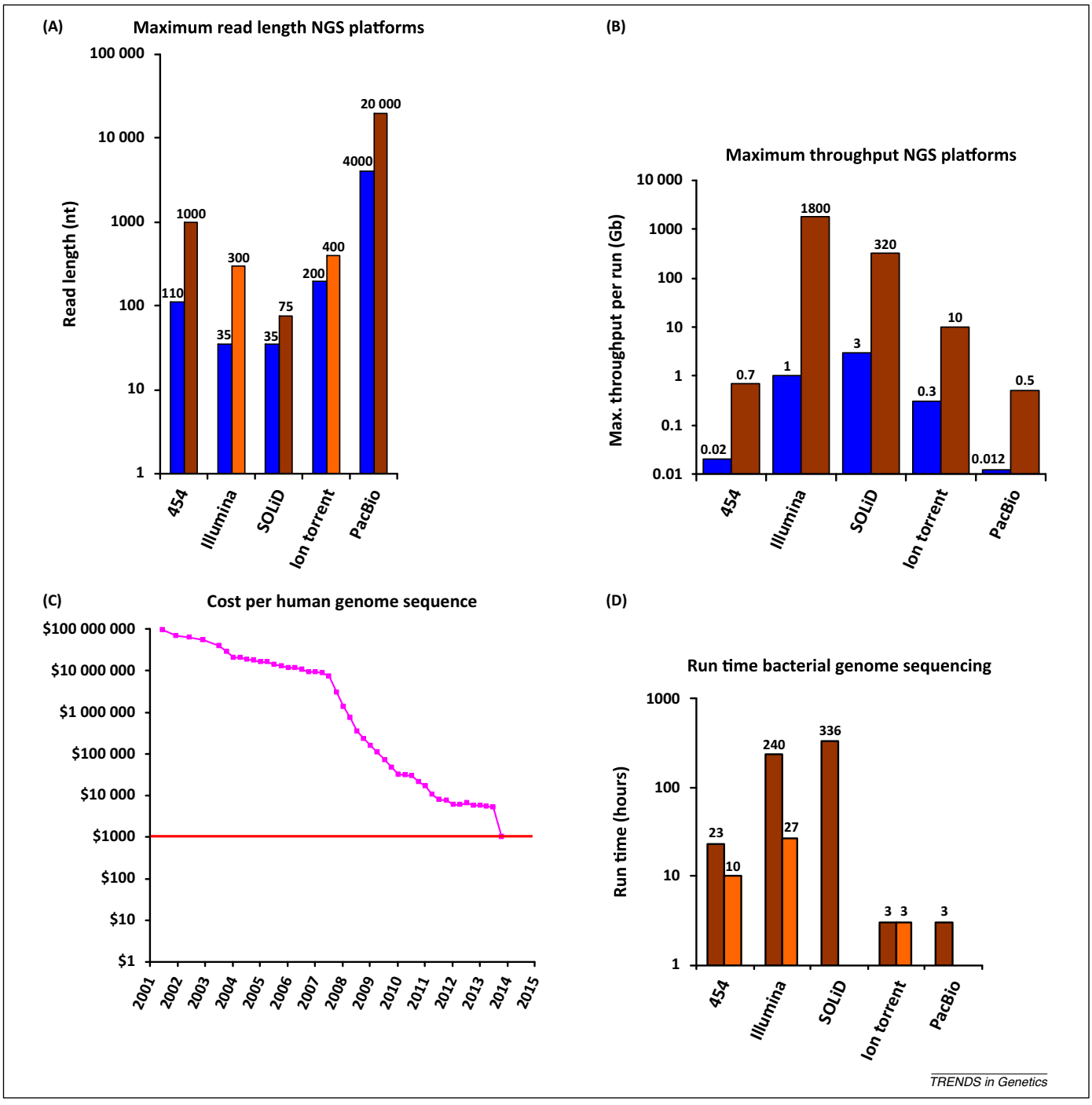

Figure 1. Evolution of high-throughput sequencing platforms. **(A)** Blue bars: maximum read length of the first commercially available sequencing instruments by 454, Illumina/Solexa, SOLiD, Ion Torrent, and Pacific Biosciences (PacBio). Orange bars indicate the maximum read length that can be obtained with these technologies today; dark orange stands for the large instruments of the different technologies, whereas light orange indicates the bench-top versions. **(B)** Maximum throughput of the first commercially available sequencing instruments (blue bars) and the current maximum throughput (dark orange bars). Note that, for a given technology, current maximum read length and throughput are not necessarily obtained with one and the same instrument. For example, Illumina's bench-top sequencer, the MiSeq, generates the longest reads but the large HiSeq X Ten has the highest throughput. **(C)** Graph showing the evolution of the cost of sequencing a human genome from 2001 until today. Costs have sharply decreased over the recent years thanks to the appearance of next-generation sequencing (NGS) technologies and their subsequent upgrades. Very recently, the milestone of the US$1000 genome has been reached with Illumina's HiSeq X Ten system. **(D)** Times to complete a typical run to sequence a bacterial genome using the large instruments of the various manufacturers (dark orange bars) or their new bench-top machines (light orange bars). For both 454 instruments, a 400 nt run was considered. For Illumina's HiSeq1000 and MiSeq, 100 nt and 150 nt paired-end runs were considered, respectively. For SOLiD's 5500W Series Genetic Analyzer, a 50 nt paired-end run was considered, and for Ion Torrent's large Proton and the bench-top PGM a 200 nt run was considered. For PacBio, run time rather than length is set, and a 3 h 'movie' is sufficient for bacterial genome sequencing. For SOLiD and PacBio, no bench-top machines are available. Although run time is an important factor for clinical applications such as bacterial genome sequencing, other important parameters include purchase pricing, costs per run, and error rates, which are not discussed here. For a comprehensive comparison of bench-top NGS platforms see [79].

of the relationship between genomic variation and phenotype [34]. In addition, WGS is being increasingly used for translational research, such as forensic genetics [35], agrigenomics (agricultural genomics) [36,37], and clinical diagnostics. An example of the latter is genetic disease diagnosis. WGS has the potential for simultaneous and comprehensive diagnostic testing of likely monogenic illnesses, which accelerates molecular diagnosis and minimizes the duration of empirical

Box 2. Pros and cons of the different NGS technologies

454
Pros. The long reads (1 kb maximum) are easier to map to a reference genome, and are an advantage for *de novo* genome assemblies or for metagenomics applications. Run times are relatively fast (~23 h)(http://www.454.com).
Cons. Relatively low throughput (about 1 million reads, 700 Mb sequence data) and high reagent cost. High error rates in homopolymer repeats [7]. But, most importantly, Roche announced that it will shut down 454 and stop supporting the platform by mid-2016 (http://www.fiercediagnostics.com/story/roche-close-454-life-sciences-it-reduces-gene-sequencing-focus/2013-10-17).
Illumina/Solexa
Pros. Illumina is currently the leader in the NGS industry and most library preparation protocols are compatible with the Illumina system. In addition, Illumina offers the highest throughput of all platforms and the lowest per-base cost [8]. Read lengths of up to 300 bp, compatible with almost all types of application.
Cons. Sample loading is technically challenging; owing to the random scattering of clusters across the flow cells library concentration must be tightly controlled. Overloading results in overlapping clusters and poor sequence quality. Another problem is the requirement for sequence complexity. Low-complexity samples such as 16S metagenomics libraries must be diluted or mixed with a reference PhiX library to generate diversity.
SOLiD
Pros. Second (after Illumina) highest throughput system on the market. The SOLiD system is widely claimed to have lower error rates, 99.94% accuracy [8], than most other systems owing to the fact that each base is read twice.
Cons. Shortest reads (75 nt maximum) of all platforms, and relatively long run times (Figure 1A,D). Less-well-suited for *de novo* genome assembly. The SOLiD system is much less widely used than the Illumina system and the panel of sample preparation kits and services is less well developed.
Ion Torrent
Pros. Semi-conductor technology, no requirement for optical scanning and fluorescent nucleotides. Fast run times; a typical run takes only a few hours. Broad range of applications.
Cons. This technology suffers from the same issue as 454 with high error rates in homopolymers.
PacBio
Pros. Extremely long reads of 20 kb and even longer make this technology an ideal tool to finish genome assemblies or to improve existing draft genomes. Another advantage is that run times are fast (typically a few hours).
Cons. High cost, US$2–17 per Mb, high overall error rates (~14%) (http://www.molecularecologist.com). Lowest throughput of all platforms (maximum ~500 Mb). Limited range of applications.

treatment [38,39]. Other important clinical applications include the sequencing of pathogenic outbreak strains and infectious disease surveillance [40].

An exciting new field is single cell genomics. A major objective of this field is to reconstruct cell lineage trees using somatic mutations that arise due to DNA replication errors. As a result, each cell in a multicellular organism carries a genomic signature that is probably unique [41]. Cell lineage trees provide important information and have applications in developmental biology [42] and tumor biology. For example, sequencing the genomic DNA of individual breast cancer cells allowed researchers to reconstruct the tumor population structure and evolutionary history [43].

For many applications it is neither practical nor necessary to sequence entire genomes, and various approaches that address defined regions of the genome have emerged. One of these is whole-exome sequencing (WES), in which only the coding regions of the genome are sequenced [44]. The exome represents less than 2% of the human genome, but contains ~85% of known disease-causing variants [45]. For this reason, WES has been extensively used for clinical studies in the recent years, and is giving rise to promising novel diagnostic tools that have the potential to transform medical healthcare in the near future. For example, an experimental approach for comprehensive WES of circulating tumor cells from cancer patients has recently been described [46].

An even higher level of targeting is achieved through amplicon sequencing, in which selected genome regions are amplified by PCR. This method is well suited for disease-targeted tests focusing on a limited number of disease-related variants. Advantages are better coverage of the relevant disease genes and reduced risk of missing important variants due to automated data filtering in exome-seq pipelines [47]. Another common amplicon application is sequencing the bacterial 16S rRNA gene across a number of species, a widely used method for studying phylogeny and taxonomy, particularly in diverse metagenomic samples [48]. This method has been used to evaluate bacterial diversity in a range of environments, allowing researchers to characterize microbiomes from samples that are otherwise difficult or impossible to study.

Instead of targeting selected regions of interest, a genome complexity reduction method that provides an unbiased sampling of a target genome has been developed. This method is based on the sequencing of restriction-site-associated DNA tags (RAD-seq) and has been designed to interrogate a number of positions scattered across a genome [49]. The level of complexity reduction depends on the restriction enzyme used, but typically a few percent of the target genome is covered, thus allowing population-scale studies of even the largest plant and animal genomes. RAD-seq can be applied not only to model organisms with well-assembled genomes but also to non-model organisms lacking complete reference genomes and is thus a great tool for SNP discovery and genotyping in those organisms [50].

RNA sequencing

Early RNA-seq studies often used protocols that did not preserve strand information. However, the eukaryotic transcriptome is much more complex than previously thought and many genes produce antisense transcripts [51]. To deal with this complexity, a number of strand-specific RNA-seq protocols have been developed, the first of which appeared in 2008 [14]. These protocols have enabled the identification of novel antisense regulatory transcripts that may have important biological functions [52–54]. Transcriptome analysis can now also be performed at the level of single cells owing to innovative sample preparation methods (see above). Unlike classical methods, which consider mixtures of heterogeneous cell populations, single cell transcriptomics provides a much more detailed view of transcription dynamics. For example, the analysis of transcriptomes from single cells has revealed that there can be substantial transcriptional heterogeneity among

Box 3. Basic principles of NGS library preparation

An elementary step in each NGS workflow is the conversion of the source nucleic acid material into a sequencing library. A wide variety of NGS library preparation protocols exist, but they all have in common the fact that (fragments of) DNA or RNA molecules are fused with platform-specific adapters. Long DNA or RNA molecules are first fragmented into a suitable size (~50–500 nt) followed by adapter addition. Next, a size selection step is usually performed to enrich further for molecules of the desired size and to eliminate free adapters. Last, PCR is usually performed to select for molecules containing adapters at both ends and to generate sufficient quantities for sequencing (Figure IA). At their extremities, the adapters usually contain elements for the immobilization of library molecules on a solid surface (emPCR beads or the surface of a glass slide, see also Box 1) (Figure IB) and amplification. Adjacent sequences serve as priming sites for sequencing. Either one or both adapters may contain sequencing priming sites. In Illumina libraries one adapter usually contains a 'read 1' sequencing primer site, which is used for single-end sequencing (in which only one end of the insert is sequenced) or for the first read of paired-end sequencing (in which both ends of the insert are sequenced). The other adapter then contains a 'read 2' primer site and a site for an index primer, which is used to read the unique index sequence, allowing libraries in a multiplexed run to be distinguished. In some library types, such as for 'dual index' sequencing, both adapters contain an index to improve the multiplexing capacity; up to 96 unique barcodes are possible with this system.

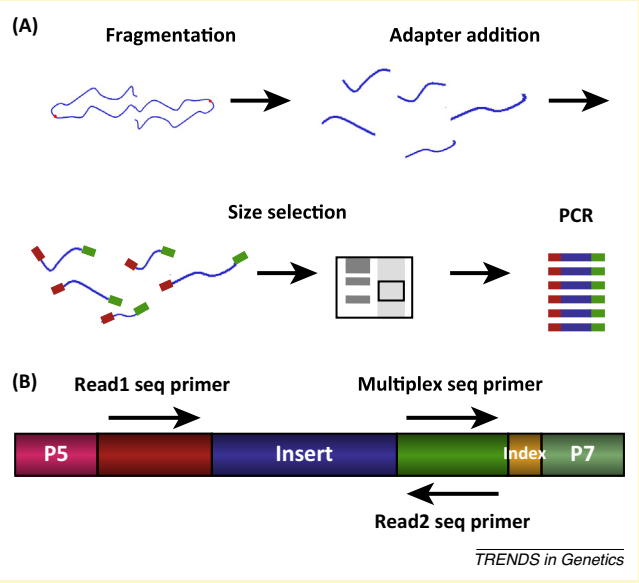

Figure I. (A) Simplified representation of a typical NGS library preparation workflow. **(B)** Architecture of a standard Illumina NGS library. Adapters are indicated in two-tone red and green, the insert is in blue. An index or barcode is indicated in yellow. Sequencing primers are indicated by arrows. Shown is an Illumina library, in which 'P5' and 'P7' are the sequences used for flow cell attachment and amplification. In other technologies the names of these terminal sequences are different but the principle is the same.

seemingly identical cells [55]. A recently published pioneering study describes a method called fluorescent *in situ* RNA sequencing (FISSEQ), which enables not only the study of the transcriptomes of single cells but also the determination of the precise location of each transcript within the cell [56].

In addition to whole-transcriptome analysis, novel methods have been developed to study specific subpopulations of transcripts or transcripts engaged in particular processes. As for DNA-seq, targeted RNA-seq approaches exist, based on capture using biotinylated oligonucleotides [57], or on PCR amplicons [58]. Capture-based RNA-seq (CaptureSeq), using tiling arrays targeting a specific portion of the transcriptome, allows sequencing that portion to a depth almost impossible with conventional RNA-seq and permits the discovery and assembly of extremely rare transcripts. Using CaptureSeq, a striking diversity of novel isoforms of well-annotated protein-coding loci was detected, demonstrating that for even well-annotated genes considerable complexity remains to be resolved [54].

A limitation of classical RNA-seq is that it measures RNA steady-state levels, which often do not directly reflect transcriptional activity or the rate of protein synthesis. Several years ago, a method was developed to visualize transcription at single-nucleotide resolution by specifically sequencing nascent transcripts. Native elongating transcript sequencing (NET-seq) is based on the immunoprecipitation of RNA polymerase (RNAP) followed by deep sequencing of the 3′ ends of co-precipitated nascent RNAs [59]. NET-seq is an alternative to RNAP ChIP-seq that provides higher resolution and retains the RNA strand information.

Another method for the specific and quantitative detection of translated RNA sequences maps ribosomes on mRNAs using nuclease footprinting followed by sequencing 28–30 nt long transcript regions shielded by ribosomes. This technique, ribo-seq, has been used to study translational control of gene expression [60], annotate translated sequences [61], and study mechanisms of protein synthesis [62,63].

Location-based techniques

Originally, ChIP-seq was developed to identify *in vivo* protein–DNA interactions [64]. It has been extensively used to study a wide diversity of biological processes and, in more-recent years, a wealth of variations on this technique has been developed. One such variation, 'ChIP-exo', localizes protein–DNA interactions at single-nucleotide resolution. In this approach, immunoprecipitated protein–DNA complexes are treated with 5′–3′ exonuclease, leaving a homogeneous 5′ border at a fixed distance of the bound protein. This allowed the determination of the precise locations of transcription pre-initiation complexes (PIC) across the yeast genome [65].

Protocols have also been designed for the genome-wide analysis of: (i) RNA–protein interactions using techniques based on crosslinking immunoprecipitation (CLIP), including CLIP-seq [66], iCLIP [67] and PAR–CLIP [68]; (ii) RNA–DNA interactions, using CHART [69] and CHiRP [70]; and (iii) DNA–DNA interactions using chromosome conformation capture (3C)-based methods, including circularized chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), Hi-C, and chromatin interaction analysis by paired-end tag sequencing

(ChIA–PET) [71]. Using these types of approaches, the three-dimensional (3D) organization of genomes can be explored at unprecedented resolution [72], which deeply changed our understanding of chromatin regulation in relation to fundamental cellular processes. For example, long-range interactions between regulatory elements and target genes have been reported throughout the genome, indicating that 3D chromatin organization represents a general higher order transcriptional control mechanism [73].

Other approaches include methods that interrogate the openness of chromatin, for example DNaseI-seq [74], or that measure DNA methylation. In the latter case, bisulfite sequencing, in which unmethylated cytosine is converted to uracil while methylated cytosine is left unchanged, has been widely used. This method offers higher resolution than other approaches, and it has been successfully used to map the complete methylomes of various human cell types and has allowed the discovery of altered methylation profiles in cancer cells [75]. Finally, even further increasing the diversity of available techniques, novel methods to detect alternative secondary structural motifs in genomic DNA or RNA have recently been developed [76,77].

Concluding remarks

The advent of NGS has enabled researchers to study biological systems at a level never before possible. As the technologies have evolved, an increasing number of sample preparation methods and data analysis tools have spawned an immense diversity of scientific applications. NGS has thus become a key technology in basic science and is rapidly becoming an established tool in translational research as well. Ongoing cost reduction and the development of standardized pipelines will probably make NGS a standard tool for more-routine applications in the near future. Clinical laboratories have already started to use NGS as a diagnostic tool, and forensic research, which still largely relies on Sanger sequencing today, is starting to use NGS. The significant increase in sample throughput would be beneficial in this field, and NGS methods may be able to solve specific cases that would not be possible with Sanger sequencing [35].

The release of Ilumina's HiSeq X Ten has reduced the cost of sequencing a human genome to US$1000 and is therefore likely to stimulate population genomics strongly. Producing tens of thousands of genomes, or so-called 'factory-scale' sequencing will revolutionize the study of population diversity and help us to understand the genetic basis of health and disease better. An enormous benefit of introducing large-scale sequencing into clinical research and ultimately healthcare is joining genomic and phenotypic information related to health and disease to find new genetic associations. This will advance our understanding of the functional consequences of DNA mutations, and improve our ability to diagnose and predict outcomes of diseases for individual patients. Eventually, this should bring personalized medicine closer to a reality.

Currently, Illumina, which offers the highest throughput and the lowest per-base cost, is the leading NGS platform. Other promising technologies are starting to appear, however. An example is Nanopore sequencing, which is based on the transit of a DNA molecule through a pore while the sequence is read out through the effect on an electric current or optical signal [78]. Nanopore is considered a third-generation technology because it enables the sequencing of single molecules in real time. Major advantages are direct sequencing of DNA or RNA molecules without the need for library preparation or sequencing reagents. Very recently the first commercially available sequencer that uses this technology, the MinION™, was pre-released by Oxford Nanopore Technologies (http://www.nature.com.gate1.inist.fr/news/data-from-pocket-sized-genome-sequencer-unveiled-1.14724). Unlike the bulky, expensive sequencing machines on the market, the MinION™ is an inexpensive hand-held device. It is capable of producing reads of up to 10 kb, but the data are still of insufficient quality owing to systematic errors and further improvement will therefore be required. Quantum Biosystems has developed an electronic single molecule DNA sequencing technology that resembles Nanopore, and has very recently published the first raw sequence data (http://www.businesswire.com/news/home/20140127005012/en/Quantum-Biosystems-Demonstrates-Reads-Quantum-Single-Molecule). At this early stage reads are short (around 20 nt) but the data already seem to be of sufficient quality to sequence the 5.4 kb genome of Phi X 174.

These novel technologies clearly need to be improved and whether or not they will shake up the sequencing industry or if Illumina will remain the leading technology in the coming years remains to be seen. However, they have the potential to launch another revolution in DNA sequencing and its applications. For example, they might be able to sequence single RNA or even protein molecules directly, vastly increasing the ease of single cell genomics.

Nevertheless, significant challenges in the implementation of NGS remain, in particular data storage and processing. In the coming years, hundreds of thousands of new human genomes will double the already impressive amount of sequence data available. With so many people's genomes sequenced, confidentiality becomes an important factor. How will this information be stored and who will have access to it? Will the individuals know every detail of their genome, or only those details pertinent to disease diagnosis or treatment? How can we prevent the possible emergence of 'genetic discrimination'? Ethical issues will definitely emerge with the communalization of personal genomes, and these issues urgently need to be addressed. In addition, more-efficient approaches to data storage and analysis are needed to keep up with the increasing speed of data production.

Acknowledgments

E.v.D., H.A., and C.T. are supported by CNRS. Y.Y. is supported by Plateforme Intégrée IMAGIF – CNRS.

References

1 Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467
2 Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74, 560–564
3 International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945

4 Schloss, J.A. (2008) How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* 26, 1113–1115
5 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
6 Valouev, A. *et al.* (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18, 1051–1063
7 Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46
8 Liu, L. *et al.* (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012, 251364
9 Ju, J. *et al.* (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19635–19640
10 Shendure, J. *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732
11 Pushkarev, D. *et al.* (2009) Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* 27, 847–850
12 Schadt, E.E. *et al.* (2010) A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240
13 Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138
14 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
15 Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680
16 Green, R.E. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444, 330–336
17 Cox-Foster, D.L. *et al.* (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318, 283–287
18 Utturkar, S.M. *et al.* (2014) Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences. *Bioinformatics*
19 Quail, M.A. *et al.* (2012) Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* 9, 10–11
20 van Dijk, E.L. *et al.* (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* 322, 12–20
21 Levin, J.Z. *et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7, 709–715
22 Caruccio, N. (2011) Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by *in vitro* transposition. *Methods Mol. Biol.* 733, 241–255
23 DeAngelis, M.M. *et al.* (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* 23, 4742–4743
24 Blainey, P.C. (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.* 37, 407–427
25 Streets, A.M. *et al.* (2014) Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* http://dx.doi.org/10.1073/pnas.1402030111
26 Hatem, A. *et al.* (2013) Benchmarking short sequence mapping tools. *BMC Bioinformatics* 14, 184
27 Zhang, W. *et al.* (2011) A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* 6, e17915
28 Li, Y. *et al.* (2013) Single nucleotide polymorphism (SNP) detection and genotype calling from massively parallel sequencing (MPS) data. *Stat. Biosci.* 5, 3–25
29 Rougemont, J. and Naef, F. (2012) Computational analysis of protein-DNA interactions from ChIP-seq data. *Methods Mol. Biol.* 786, 263–273
30 Kvam, V.M. *et al.* (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* 99, 248–256
31 Hansen, K.D. *et al.* (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38, e131
32 Abecasis, G.R. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073
33 Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* 100, 659–674
34 Kilpinen, H. and Barrett, J.C. (2013) How next-generation sequencing is transforming complex disease genetics. *Trends Genet.* 29, 23–30
35 Weber-Lehmann, J. *et al.* (2014) Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Sci. Int. Genet.* 9, 42–46
36 Poland, J.A. *et al.* (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7, e32253
37 Goddard, M.E. and Hayes, B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10, 381–391
38 Kingsmore, S.F. and Saunders, C.J. (2011) Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci. Transl. Med.* 3, 87ps23
39 Saunders, C.J. *et al.* (2012) Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* 4, 154ra135
40 Lipkin, W.I. (2013) The changing face of pathogen discovery and surveillance. *Nat. Rev. Microbiol.* 11, 133–141
41 Frumkin, D. *et al.* (2005) Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* 1, e50
42 Reizel, Y. *et al.* (2011) Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLoS Genet.* 7, e1002192
43 Navin, N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94
44 Hodges, E. *et al.* (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527
45 Choi, M. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19096–19101
46 Lohr, J.G. *et al.* (2014) Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.* 32, 479–484
47 Rehm, H.L. (2013) Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* 14, 295–300
48 Faust, K. and Raes, J. (2012) Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550
49 Baird, N.A. *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376
50 Davey, J.W. and Blaxter, M.L. (2010) RADSeq: next-generation population genetics. *Brief. Funct. Genomics* 9, 416–423
51 Jensen, T.H. *et al.* (2013) Dealing with pervasive transcription. *Mol. Cell* 52, 473–484
52 van Dijk, E.L. *et al.* (2011) XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* 475, 114–117
53 Mills, J.D. *et al.* (2013) Strand-specific RNA-seq provides greater resolution of transcriptome profiling. *Curr. Genomics* 14, 173–181
54 Siegel, T.N. *et al.* (2014) Strand-specific RNA-seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. *BMC Genomics* 15, 150
55 Shalek, A.K. *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 509, 363–369
56 Lee, J.H. *et al.* (2014) Highly multiplexed subcellular RNA sequencing in situ. *Science* 343, 1360–1363
57 Mercer, T.R. *et al.* (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30, 99–104
58 Blomquist, T.M. *et al.* (2013) Targeted RNA-sequencing with competitive multiplex-PCR amplicon libraries. *PLoS ONE* 8, e79120
59 Churchman, L.S. and Weissman, J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469, 368–373
60 Guo, H. *et al.* (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835–840
61 Brar, G.A. *et al.* (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335, 552–557
62 Li, G.W. *et al.* (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 538–541
63 Stadler, M. and Fire, A. (2011) Wobble base-pairing slows *in vivo* translation elongation in metazoans. *RNA* 17, 2063–2073
64 Johnson, D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 316, 1497–1502

65 Rhee, H.S. and Pugh, B.F. (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* 21, 10
66 Sanford, J.R. *et al.* (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.* 19, 381–394
67 Konig, J. *et al.* (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909–915
68 Hafner, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141
69 Simon, M.D. *et al.* (2011) The genomic binding sites of a noncoding RNA. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20497–20502
70 Chu, C. *et al.* (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* 44, 667–678
71 Dekker, J. *et al.* (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390–403
72 Duan, Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature* 465, 363–367
73 Ethier, S.D. *et al.* (2012) Discovering genome regulation with 3C and 3C-related technologies. *Biochim. Biophys. Acta* 1819, 401–410
74 Hesselberth, J.R. *et al.* (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods* 6, 283–289
75 Lee, E.J. *et al.* (2013) Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer Lett.* 340, 171–178
76 Lam, E.Y. *et al.* (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* 4, 1796
77 Ding, Y. *et al.* (2014) *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505, 696–700
78 Clarke, J. *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4, 265–270
79 Loman, N.J. *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439