Biochimica et Biophysica Acta 1842 (2014) 1932–1941

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbadis

Review

Next generation sequencing technology: Advances and applications ☆

H.P.J. Buermans [1], J.T. den Dunnen *

Leiden Genome Technology Center, Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 22 November 2013
Received in revised form 5 June 2014
Accepted 15 June 2014
Available online 1 July 2014

Keywords:
Next generation sequencing
Sequence by synthesis
Nanopore
Single molecule sequencing
Basic technology
Applications

ABSTRACT

Impressive progress has been made in the field of Next Generation Sequencing (NGS). Through advancements in the fields of molecular biology and technical engineering, parallelization of the sequencing reaction has profoundly increased the total number of produced sequence reads per run. Current sequencing platforms allow for a previously unprecedented view into complex mixtures of RNA and DNA samples. NGS is currently evolving into a molecular microscope finding its way into virtually every fields of biomedical research. In this chapter we review the technical background of the different commercially available NGS platforms with respect to template generation and the sequencing reaction and take a small step towards what the upcoming NGS technologies will bring. We close with an overview of different implementations of NGS into biomedical research. This article is part of a Special Issue entitled: From Genome to Function.

1. Introduction

The growing power and reducing cost sparked an enormous range of applications of Next generation sequencing (NGS) technology. Gradually, sequencing is starting to become the standard technology to apply, certainly at the first step where the main question is "what's all involved", "what's the basis". It should be realized that for many applications sequencing would always have been the method of choice, yet it was science-fiction, technically unthinkable and later possible but far too costly. We perform genome-wide association studies (GWAS) using SNP-arrays simply because we cannot afford to perform whole-genome sequencing in ten-thousands of individuals. This is changing rapidly and sequencing will become our molecular microscope, the tool to get a first look. Although replication, transcription, translation, methylation and nuclear DNA folding are completely different processes, they can all be studied using sequencing.

An important advantage of sequence data is its quality, robustness and low noise. It should be noted that a successful NGS project requires expertise both at the wet lab as well as the bioinformatics side in order to warrant high quality data and data interpretation. The sequence itself is hard evidence of its correctness. A sequencing system will not produce "random" sequences and when it does this becomes evident immediately from QC calls obtained from spike-in controls. Furthermore random sequences will have no match and can be easily discarded during data analysis and when their number exceeds a certain threshold it is evident that there is a serious problem somewhere in the study.

2. Sequence library preparation

All currently available sequencing platforms require some level of DNA pre-processing into a library suitable for sequencing. In general, these steps involve shearing of high molecular weight DNA into an appropriate platform-specific size range, followed by an end polishing step to generate blunt ended DNA fragments. Specific adapters are ligated to these fragments by either A/T overhang or direct blunt ligation. A functional library requires having specific adapter sequences to be added to the 3′ and 5′ ends. Each of the sequence platforms uses a different set of unique adapter sequences to be compatible with the further steps of the process (Fig. 1).

Following adapter ligation Life Technologies (Solid, PGM, Proton) libraries require a nick translation step to get functional molecules while for the other technologies the sample is in principle ready for loading immediately after ligation. One may then choose to sequence these libraries directly as amplification free libraries or introduce a pre-amplification step prior to sequencing. It is important to realize that any step during pre-processing which involves amplification of the molecules [1] or which has been shown to be sequence biased, like ligations [2], will impose a selection on molecules that end up in the sequenceable libraries.

3. Current sequencing technology

The different sequence platform vendors have devised different strategies to prepare the sequence libraries into suitable templates as

☆ This article is part of a Special Issue entitled: From Genome to Function.

* Corresponding author: Tel.: + 31 715269400; fax: + 31 71 5268285.
E-mail addresses: h.buermans@lumc.nl (H.P.J. Buermans), ddunnen@humgen.nl (J.T. den Dunnen).
[1] Tel.: +31 715269400; fax: +31 71 5268285.

http://dx.doi.org/10.1016/j.bbadis.2014.06.015
0925-4439/

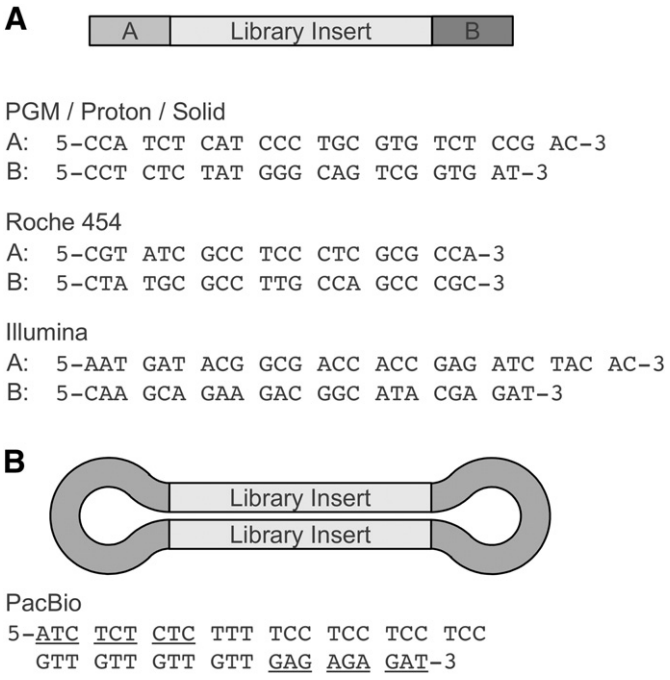

Fig. 1. Structure of sequence library molecules for the different technologies. Linear library molecules (Panel A) contain different adapter sequences at the 5′ [A] and 3′ [B] ends of the library inserts. Circular library molecules (Panel B) contain identical adapter molecules at both ends of the insert.

well as to detect the signal and ultimately read the DNA sequence. For the Illumina, Solid, PGM and 454 systems a local clonal amplification of the initial template molecules into polonies [3] is required to increase the signal-to-noise ratio because the systems are not sensitive enough to detect the extension of one base at the individual DNA template molecule level. On the other hand, the Heliscope and PacBio SMRT systems do not need any pre-amplification steps as these systems are sensitive enough to detect individual single molecule template extensions. The different strategies to generate the sequence reads also lead to differences in the output capacity for the different platforms (Table 1). Below we will focus on the newer sequencing platforms, being the Illumina, LifeTechnologies Semiconductor sequencing and PacBio. Other older platforms will briefly be discussed in Online Supplement 1.

3.1. Illumina technology

All of the enzymatic processes and imaging steps of the Illumina technology take place in a flow cell. Depending on the specific Illumina platform it may be partitioned into 1 (miSeq), 2 (HiSeq2500) or 8 (HiSeq2000, HiSeq2500) separate lanes. The Illumina platform uses bridge amplification for polony generation and a sequencing by synthesis (SBS) approach (Fig. 2A). Forward and reverse oligos for amplification (one with a cleavable site), complementary to the adapter sequences introduced during the library preparation steps, are attached to the entire inside surface of the flow cell lanes. The first step for loading the library onto the flow-cell is denaturation of the dsDNA fragments into individual ssDNA molecules. When on the flow-cell, these hybridize to the oligo nucleotides on the surface (Fig. 1A; step 1) which are used as primers to form an initial copy of the individual sequencing template molecule (Fig. 1A; step 2). The initial library molecules are removed and the copied, flow cell-attached fragments are used to generate a cluster of identical template molecules using isothermal amplification. This is done through cyclic alternations of three specific buffers that mediate the denaturation, annealing and extension steps at 60 °C. During these steps the 3′ end of the copied library molecules can hybridize to the complementary oligos on the flow cell, thus forming a bridge structure (Fig. 1A; steps 3–5). The final step is to remove one strand of the dsDNA fragments using the cleavable site in the surface oligo (Fig. 1A; step 6) and to block all 3′ ends with ddNTP to prevent the otherwise open 3′ ends to act as sequencing primer sites on adjacent library molecules [4].

With optimal loading of library molecules one flow-cell lane will yield approximately 800–1000 K clusters per mm^2. Optimal amounts depend not only on the concentration of the library, but also on the length of the molecules. Short molecules yield clusters with a small area that are denser and therefore generate more intense signals. Loading a wide fragment size distribution will generate clusters varying widely in size and signal strength which may impair the number of passing filter reads.

Bridge amplification is not a very efficient method for clonal amplification, i.e., the 35 cycles of isothermal amplification yield a mere ~1000 copies of the initial molecule. Moreover, there will be predominantly outward growth of the clusters, there is a high probability of the template strands to re-hybridize instead of annealing to a new primer site on the glass surface and there is both an upper and a lower limit to the length of the template molecules that can be reliably amplified. In addition, DNA polymerases, which are known to have biases towards specific DNA templates are used during the amplification processes. The bridge amplification scheme that Illumina exploits yields a high number of clusters, i.e., with good loading of the flow cell, the total number of reads generated per Hiseq2000 lane may reach ~180 million. With a paired-end 2 × 100 bp read format the total output of one flow-cell lane is up to ~36 Gb. A full run of 2 flow cells sequencing in parallel may yield ~600 Gb of data.

During sequencing, the polonies on the flow cell are read one nucleotide at a time in repetitive cycles. During these cycles, fluorescently labeled dNTPs are incorporated into the growing DNA chain. Each of the four dNTP species (A, C, T, G) has a single different fluorescent label which serves to identify the base and act as a reversible terminator to prevent multiple extension events. After imaging the fluorescent group is cleaved off, the reversible terminator is de-activated and the template strands are ready for the next incorporation cycle. The sequence is read by following the fluorescent signal per extension step for each cluster. Under ideal circumstances, all bases within a cluster will be extended in phase. However, a small portion of the molecules do not extend properly and fall either behind (phasing) or advance a base (pre-phasing). Over many cycles, these errors will accumulate and decrease the signal to noise ratio per cluster, causing a decrease in quality towards the ends of the reads.

The cycle time for the HiSeq2000 is approximately 1 h. The major contributor is the imaging of the flow-cell. The enzymatic reactions take very little time at all. By reducing the imaging time, the whole sequencing process can be sped up considerably. This is implemented in the miSeq and Hiseq2500 platforms by providing the option to decrease the total surface area to be imaged. In rapid mode, cycle time can thereby be reduced to 5 and 10 min for the miSeq and Hiseq 2500, respectively. Furthermore, with optimized reagent kits for these short cycle times it is possible to achieve a 2 × 300 bp paired end run on the miSeq, with 85% of data points above Q30 and run times of ~65 h. However, the increased sequencing speed does come at a price. With the decreased surface area, the total number of data points that can be generated per run will reduce, increasing sequencing cost per nucleotide significantly.

Early 2014, Illumina has announced the release of two new sequencer models, i.e., the NextSeq 500 and the HiSeq X Ten. The former system was designed to be a highly flexible, smaller version of the HiSeq2500, providing both medium (40 Gb) and a high output (120 Gb) modes both with run times under 30 h. The HiSeq X Ten was designed for one main purpose: enabling whole human genome sequencing and reaching the $1000 genome in run costs. The main advancement enabling this is the introduction of the patterned flowcells. In contrast to the spatial random cluster generation of the HiSeq and MiSeq flowcells, the X Ten flowcells contain a pre-formatted grid of nano-wells, which each can produce

Table 1
Summary of the output per sequencing technology platforms.

		Sequence by	Detection	Run types	Run time	Read length (bp)	# reads per run	Output per run	Remarks
Roche	GS FLX Titanium XL+	Synthesis	Pyrophosphate detection	Single end	23 h	700	1 million	700 Mb	
	GS Junior System	Synthesis	Pyrophosphate detection	Single end	10 h	400	0.1 million	40 Mb	
LifeTechnologies	Ion torrent	Synthesis	Proton release	Single end	4 h	200–400	4 million	1.5–2 Gb	Ion318 Chip
	Proton	Synthesis	Proton release	Single end	4 h	125	60–80 million	8–10 Gb	IonP1 chip
	Abi/solid	Ligation	Fluorescence detection of di-base probes	Single & paired-end	10 days	75 + 35	2.7 billion	300 Gb	
Illumina/solexa	HiSeq2000/2500	Synthesis	Fluorescence; reversible terminators	Single & paired-end	12 days	2 × 100	3 billion	600 Gb	High output mode
	MiSeq	Synthesis	Fluorescence; reversible terminators	Single & paired-end	65 h	2 × 300	25 million	15 Gb	
Pacific biosciences	RSII	Single molecule synthesis	Fluorescence; terminally phospholinked	Single end	2 days	50% of reads > 10 kb	0.8 million	5 Gb	16 SMRT cells
Helicos	Heliscope	Single molecule synthesis	Fluorescence; virtual terminator	Single end	10 days	~30	500 million	15 Gb	Two flowcells in parallel

one sequence polony. This allows for optimized cluster densities and in combination with faster scanning protocols, specifically tailored software and new sequencing chemistry this system can produce ~600 Gb of data in a single day or ~1.8 terabases within three days, which is enough for 5 and 15 whole human genomes, respectively.

3.2. Ion torrent technology

For the PGM/Proton sequence platforms the sequence templates are generated on a bead or sphere via emulsion PCR (emPCR) [5,6]. An oil–water emulsion is created to partition small reaction vesicles that each ideally contains one sphere, one library molecule and all the reagents needed for amplification. Two primers that are complementary to the sequence library adapters are present, but one is only present in solution while the other is bound to the sphere. This serves to select for the library molecules with both an A and a B adapter while excluding those molecules with two A or B adapters from loading on the beads during emPCR (Fig. 2B). In addition, this ensures a uniform orientation of the sequence library molecules on the sphere. During the emPCR steps, individual library molecules get amplified to millions of identical copies that are bound to the beads to allow ultimate detection of the signal.

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

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

The dNTPs are added in a predefined flow order. At the first release of the system, this order was a repetitive T–A–C–G sequence. As with Illumina sequencing, not all of the template molecules on a templated sphere get extended in perfect synchrony. On average 0.5–1% of the molecules deviate from the flow either because they lag behind due to improper extension or they advance ahead due to carry over of dNTPs from a previous cycle. In order to minimize this de-phasing, the flow order was changed to a more sophisticated sequence that incorporated A–T–A catch-up type flows. This scheme allows incomplete extension of

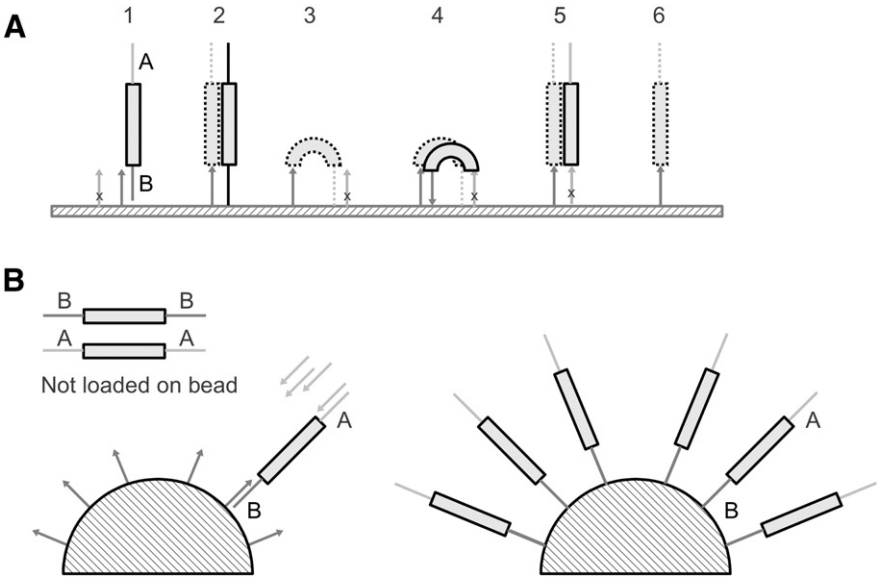

Fig. 2. Template generation via bridge amplification (A) or emPCR (B).

the A nucleotide to catch up after the T base. Although this does come at a cost of decreased overall read length, the overall quality of the read does improve. Still, the quality of the reads gradually decreases towards the ends of the reads. By taking into account the flow order, it is possible to make flow-aware base caller algorithms and flow-space aware aligner software and variant detection tools that take the actual flow order into account when processing the data in order to generate higher accuracy data [7,8]. The present error rate for substitutions is ~0.1% [8] which is very similar to that of the Illumina systems. The main point of criticism the system endures is the homo-polymer errors. Despite many improvements the 5-mer homo-polymer error rate is still at 3.5% [8].

Since the initial release of the Ion torrent platform, this technology has evolved at a very rapid pace. The output specs of the first Ion-314 chip were a mere 10 Mb. Through increasing the total surface area of the chips and the sensor well density, all on the 350 nm CMOS technology, in addition to increasing the average read length from 100 up to 400 bp, the newest Ion-318 chips produce ~1 Gb. For the Ion Proton System 110 nm CMOS technology was used to manufacture the Proton-I chips. The diameter of the spheres and the sensor wells decreased which allowed the number of wells to increase to ~165 million per chip. The Proton-I chips currently yield 60–80 million reads per run, reaching 10 Gb. This is enough to sequence two human exomes at ~50× coverage. The announced Proton-II chip will have 4× the number of sensor wells, with an expected output of ~32 Gb per chip promising to generate a whole human genome at ~10× coverage, still within the 4 h run time. This output puts it at par with a paired-end run on single HiSeq2000 lane.

Life Technologies have developed an alternative method to generate polonies called Wildfire [9]. The process generates clusters on a solid surface using isothermal amplification without denaturation or amplification cycles. Although initially designed for the Solid 5500 system, it is likely that this method could be applied to the Ion torrent semiconductor sequencing as well. This may involve spheres as an intermediate carrier or clusters may be generated directly into the sensor wells. A Proton-III chip has been announced that will double the number of wells to 1.2 billion, leading to an expected output of ~64 Gb per run. With these output levels, the Ion Proton will become a competitor to the current Illumina HiSeq systems.

3.3. Pacific biosciences technology

The principles underlying the pacific biosciences single molecule real-time (SMRT) sequencing technology are quite different from those of the above mentioned sequencers. First of all, the technology works with single molecule detection, i.e., the optics used are sensitive enough to detect incorporation of one fluorescently labeled nucleotide. Consequently, template preparation does not require any amplification steps, and the prepared library molecule is the sequencing template. Library preparation is similar to shotgun libraries for the other platforms, i.e., fragmentation of the gDNA to the required size, here multi-kb size, followed by end repair and either A/T overhang or direct blunt adapter ligation. The main difference is that the adapters have a hairpin structure (SMRT loop adapters) so that after ligation the dsDNA fragments will have become circular (Fig. 1B). Pre-preparation of the sequencing template consists of annealing a sequencing primer to the ssDNA region of the SMRT loop adapters, followed by binding of the DNA polymerase to form the active polymerization complex. A combination of insert size and read length will determine whether a short molecule is read several times (CCS or consensus circular sequencing) or a long molecule only once.

The standard PacBio DNA library preparation starting amounts can be as high as 1–5 μg total genomic DNA or 0.5–1.0 μg sheared and size selected DNA fragments for 10 kb libraries. These numbers are unexpectedly high for a system that reads single DNA molecules. These high input requirements may limit the use of the PacBio system for low input applications such as ChIP-Seq or single-cell genomics. The bottleneck lies at the stringent XP bead clean-up steps and Exo III and VII treatments during library preparation to exclude short and/or non-circular fragments from the final library pool. A lot of sample is lost during these steps, but not taking these measures which would lead to suboptimal output of the runs. Alternative methods have been explored that bypass these stringent library preparation steps and immediately proceed to sequencing the unprocessed DNA template molecules either by specifically primed or random hexamer primed sequencing on plasmid DNA. Although the output of the SMRT cells was far below standard specs, Coupland et al. were able to sequence the entire M13mp18 genome at average 56× coverage using as little as 3.1 ng input DNA [10]. These experiments demonstrated that it is possible to directly sequence small single- and double-stranded DNA genomes without the need for any DNA hungry library preparation steps.

The sequencing reaction takes place at the bottom of the ~150,000 zero-mode waveguide (ZMW) wells [11] on a SMRT cell. These ZMW are small reaction wells that each ideally contains one complex consisting of template molecule, sequencing primer and DNA polymerase bound to the bottom of the ZMW [12]. Unlike the Illumina and Life Technologies

platforms, PacBio does not rely on interrupted cycles of extension and imaging to read the template strand. Instead, the fluorescent signals of the extended nucleotides are recorded in real time at 75 frames per second for the individual ZMWs. This is achieved by a powerful optical system that illuminates the individual ZMWs with red and green laser beamlets from the bottom of the SMRT cell and a parallel confocal recording system to detect the signal from the fluorescent nucleotides [13]. The width of the ZMWs is chosen in relation to the laser wavelength such that the light cannot pass through the ZMWs but a zeptoliter sized illumination zone is formed at the bottom of the ZMWs where the active polymerase complex is bound.

As with the Illumina platform, the four nucleotide species are labeled with a different fluorescent label. However, a crucial difference is that the label for the PacBio system is terminally phospholinked nucleotides, meaning the label is cleaved off during strand extension. In addition, the nucleotides do not contain a terminator group allowing continuous extension of the growing DNA copy. When a nucleotide complementary to the template is bound in position by the polymerase within the illumination zone of the ZMW, the identity of the nucleotide is recorded by its fluorescent label. During extension the label is cleaved off, diffuses outside the illumination zone, and the complex is ready for the next extension. In essence, the system records a movie of the activity of the polymerase during a rolling circle amplification of the template. The polymerase used for sequencing in the PacBio system is a modified version of phi29 [14] which although exhibiting reduced 3–5 exonuclease activity, still has many of the properties of the original phi29 like high processivity of several hundred kilobases, low error rate of ~10e − 5 [15], no GC bias and strand displacement properties [16].

Through several advancements at the technological, bioinformatics and chemistry level, the average output per SMRT cell has increased considerably since the release of the system early 2011. The output has mainly increased though achieving longer average read lengths, starting from the ~1 kb at the time of release, while the number of reads passing filter remained ~50–60 K per SMRT cell. Although the maximal achievable read lengths that can be obtained with this technology are directly related to the length of the sequencing time, not all polymerase complexes reach identical read lengths. The main cause for this is photo-damage of the phi29 polymerase which terminates the sequencing reaction. In the C3 chemistry, photo-protected nucleotide analogs are used which shield the polymerase from damage and consequently, half of the reads are over 8–10 kb in length when using the P5 enzyme (Fig. 3). With the current max. movie length of 180 min, read lengths of over 40,000 bases have been reported. With these specifications the output per SMRT cell reaches ~400 Mb for the PacBio RSII.

One much debated aspect of the PacBio data is the high single pass error rate of 10–15%, the majority of which are insertion/deletion errors with only a small fraction of miscalls. It is important to keep in mind that in contrast to the other sequencing technologies, the errors in the PacBio reads are randomly distributed and do not occur more frequently towards the end of the reads. This property can be used to create consensus calls from information of multiple reads covering each reference position. Consensus accuracy can reach over 99.999% by using the PacBio Quiver software [17]. Although this does require a coverage of ~40× per base, these levels can easily be reached with the output of a few SMRT cells for small (bacterial) genomes.

Despite the relatively low output of the system per SMRT cell, the PacBio long read data, absence of GC bias and insight into the kinetic state of the polymerase during sequencing present a niche of specific applications for this system that cannot be covered by any of the other currently available sequence platforms.

4. Future sequencing technology

The advancements made on sequencing technology over the last years have been impressive. However, the ultimate sequencing platform would work on single DNA or RNA molecules without any (pre-) amplification, without use of optical steps, reads of Mb to Gb in length, no GC bias, high read accuracy and would be flexible enough to generate as many sequence reads as are necessary for the specific research question at hand. In addition, it should be both cheap to acquire and run, easy to operate, have short run times and simple or no library pre-preparation steps. Needless to say, this sequencing platform does not exist, yet. In the next section we will discuss one emerging sequencing platform which may have the potential to make the next step towards these ultimate sequencers.

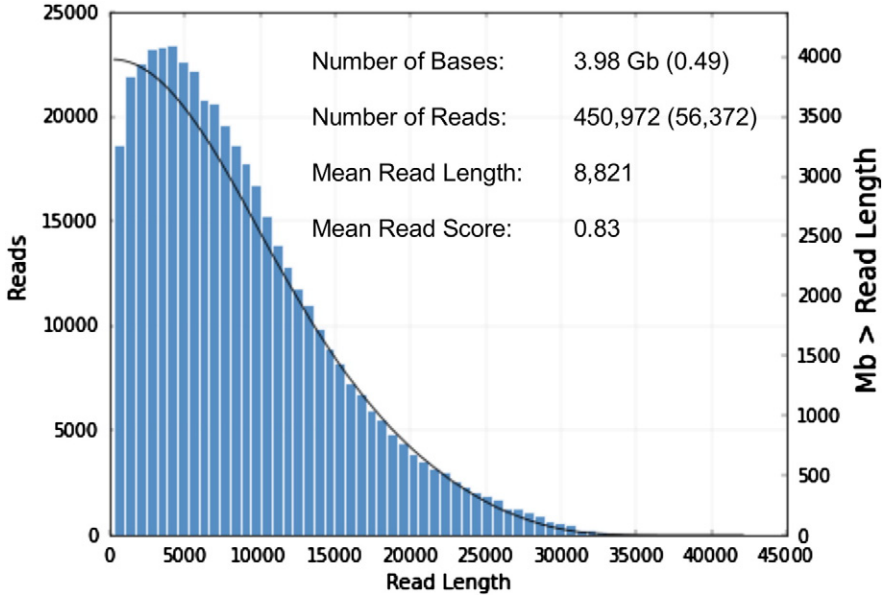

Fig. 3. Total read length distribution for PacBio reads obtained with the P5 enzyme in combination with the C3 chemistry. Blue bars (left y-axis) represent the number of reads and the black line (right y-axis) represents the total amount of data from reads longer than the read length on the x-axis. The data presented is from 8 SMRT cells run on the same library. The total number of bases and reads is indicated with the average per SMRT cell in parentheses.

4.1. Oxford nanopore sequencing

Expected characteristics for the nanopore sequencers are single molecule, amplification free, base detection without labels, long reads, low GC bias and scalable in data output. The basic principle behind the technology is tunneling of (polymer)molecules through a pore that separates two compartments. Physical presence of the molecule passing through the pore causes a characteristic temporary change in the potential between the two compartments which allows for identification of the specific molecule [18]. Two version of nanopore DNA sequencing are being developed, i.e., using the natural pore forming protein alpha-hemolysin [19] or manufactured solid state pores [20,21].

Oxford nanopore technologies (ONT) is one of the companies working on building nano-pore sequencing devices. Although ONT has put their focus on nucleic acid sequencing, in principle this technology could be applied to any (bio-)polymer as long as the molecules yield distinguishable changes in the current between the compartments. In Q4 2012 Oxford nanopore announced the early access release of their MinION system. This is a palm sized sequencing device that facilitates real-time analysis of single molecules such as DNA and RNA. However, specifics on read length, accuracy and run times are difficult to obtain.

5. NGS applications

Below we will briefly describe a range of NGS applications. Each application requires a minimal amount of reads of a pre-defined length in order to obtain a dataset which allows the researcher to draw reliable conclusions. Specifications differ for each application but in Table 2 we give a rough indication of the minimal dataset required. Still, depending on the specific research question, the number of reads and/or read length may differ.

5.1. Expression analysis

The development of micro-array technology revolutionized biomedical research, for the first time allowing overall characterization (genome-wide) of gene expression. Yet the technology had several flaws limiting sensitivity and specificity, that are overcome by sequencing. A first application of NGS was in gene expression profiling. While detection of low expressed genes using micro-arrays is limited by background noise, sensitivity of sequence based studies is predominantly limited by the depth of sequencing. Recent publications show that RNA-based gene expression can be measured in many ways. Initially, cost-directed, SAGE-like approaches were popular [22], but dropping costs now drives analysis towards full RNA sequencing (RNAseq) [23], requiring much deeper sequencing yet through analysis of the entire transcript revealing the complete picture incl. differential splicing and allelic expression. A combination of SNP-array based GWAS studies and NGS-based RNAseq is now used to study gene activity per allele and detect variants that affect transcription both in cis and trans (eQTLs) [24]. NGS has also been extensively used to study RNA structure in detail and specific methods were developed to characterize e.g. transcription initiation using the 5′-start (cap site) [25] or the 3′-end of transcripts [26].

Small RNAs were re-discovered and large studies were performed measuring e.g. miRNA expression in great detail [27,28]. Enhancers were recently found to initiate RNA polymerase II transcription, producing so-called eRNAs. Through the presence of balanced bidirectional capped transcripts (from CAGE analysis), the FANTOM5 consortium was able to obtain a genome-wide atlas of active enhancers across human cell types and tissues [29].

A recent addition is ribosome profiling [30] where not all RNA is isolated but only that RNA which at the time of sampling is being translated into protein. Besides giving a picture much closer to protein levels the technology can also be used to annotate the genome (translation initiation site, termination site and reading frame used). Finally transcripts reveal whether they contain the so-called untranslated open reading frames (uORF), in yeast often found in stress-related genes.

5.2. ChIP-seq

Another application boosted by NGS technology was the study of protein binding sites in genomic DNA, especially that of transcription factor binding sites based on Chromatin ImmunoPrecipitation (ChIP) [31]. Initially ChIP-enriched sequences were analyzed using array technology but since array content was limited choices had to be made regarding the sequences to analyze. In general promoter regions of genes and CpG-rich regions were selected. Since it was often difficult to enrich the sequences of interest to high purity, the signals obtained were often rather weak. The advantage of sequencing the ChIP-products enriched is that all sequences bound are identified and that the number of sequence read can be used to determine the sensitivity; even when enrichment is weak, increasing the number of reads can help to obtain a good signal. Many studies have been performed using ChIP-seq and DNase I hypersensitive site (DHS) mapping of which those of the ENCODE [32] and FANTOM5 [29] projects are most impressive, revealing genome wide profiles and binding sites for a range of DNA binding proteins.

5.3. Methylation

Like ChIP-seq, the power of NGS gave a boost to the study of (genome-wide) DNA methylation. Irrespective of the technology used to isolate the methylated sequences, e.g., methylated DNA immunoprecipitation-sequencing (MeDIPseq), methylated CpG island recovery assay (MIRA-Seq) or whole-genome shotgun bisulphite sequencing (WGSBS) (reviewed in Liard et al [33]), NGS-analysis clearly reveals all sequences enriched. Besides enriching methylated DNA sequences, chemical tricks can be used to identify methylated nucleotides. Sequencing both untreated and bisulphite-treated DNA will highlight the C-nucleotides that are methylated and not chemically converted resulting in a T when sequenced. Although powerful, the lower complexity of the bisulphite treated DNA (most C's converted to a T) will cause some problems during analysis (determining the origin of the reads obtained). Another complicating factor is that, since methylation is usually not 100%, a significant sequence depth will be required to get meaningful results. Single molecule sequencing technologies, like that of pacific biosciences (see above), are attractive alternatives especially since they are able to detect all DNA modifications present and not only methylation of C-residues. Many NGS methylation studies have been presented among which those of the ENCODE consortium [32]. An elegant smaller study is that by Herb et al. [34] in honeybees where they show that substantial differences exist in the methylation pattern in the brain between nurses and foragers but not between a queen and the worker castes. Interestingly, the difference between nurses and foragers was reversible.

5.4. De novo genome sequencing

A main application of NGS technology is the complete characterization of the entire genome of a particular species. After tackling first the main model organisms (yeast, *Escherichia coli*, Drosophila, Arabidopsis, mouse) and the human genome, an ever increasing number of genome sequences is currently being performed. An important player here is BGI (Beijing Genomics Institute) with statements like “Does it taste good, we'll sequence it”, “Does it look cute, we'll sequence it” and their “1000 Plant and Animal Reference Genomes Project” and Ten Thousand Microbial Genomes Project. Considerable effort has also been invested in the analysis of genomes from endangered or even extinct species, like the panda [35], mammoth [36] and early humans [37]. Together these projects produce an enormous amount of data that, besides for some evolutionary studies, remains largely unstudied.

Table 2
Recommendation for data requirements for a selection of NGS applications.

Application	# reads/sample	Run type	# read length (bp)	Remark
Transcriptome analysis				
Tag based (SAGE/CAGE)	>10 million	Single end	20–50	
SmallRNA	>10 million	Single end	20–50	
mRNA Seq	>30 million	Paired-end	>50	Efficient exclusion of rRNA derived sequences increases the resolution of the transcripts of interest
Ribosome profiling	>20 million	Single end	20–50	
ChIP-Seq	>20 million	Single or Paired-end	≥50	Specificity of the ChIP enzyme determines the # reads needed. Low specificity ~ more background = more reads needed
De novo sequencing	30× genome coverage, preferably more.	long single-end reads and Paired-end	As long as possible	Ideal PacBio long reads. Or combination of paired-end, mate-pair and PacBio.
Meta-genomics				
Tag based (ITS, 16S)	>100,000	Paired-end, long single-end reads	As long as possible	Complexity of the specific biosphere determines both the primer pairs and/or #reads per sample. Longer reads allow for better differentiation between related species
Shotgun	>100 million	Paired-end, long single reads	As long as possible	Complexity of the specific biosphere determines the library insert size and/or #reads per sample.
Methylation analysis				
Whole genome	>400 million	Paired-end	≥100	Ideal situation: all PacBio long reads on native/ unmodified shotgun libraries.
Enrichment strategies	>50 million	Paired-end	≥100	
Infections	>25 million	Single or Paired-end	≥100	~2% of cell-free DNA from plasma is of non-human origin
Non-invasive prenatal testing	>10–20 million	Single-end	>50	Trisomy detection from cell-free fetal DNA in maternal plasma
Disease gene identification diagnostics				
Whole genome	1 billion	Paired-end	≥100	30× average coverage
Exome (50 Mb)	>60 million	Paired-end	≥100	50× average coverage, >75% on target

5.5. Metagenomics

A complex variant of de novo genome sequencing is "sequence it all", metagenomics. Using brute force sequencing, simply reading all DNA sequences present in a sample, metagenomics is a way to make an inventory of what is present in a sample, of what is living where. A simple but effective application of this is trying to detect the cause of an infectious disease. Simply analyzing all DNA from control versus infected (diseased) individuals will reveal the "extra" DNA which most likely derives from the infectious agent. The approach was used successfully to identify e.g. colony collapse disorder killing honeybees [38] but also to identify the cause of diseases that killed thousands of humans in the past, inclusive the black death [39].

Metagenomics can be performed by a sequence it all approach or by focusing on specific uniformly conserved sequences like e.g. ribosomal RNA genes only. The latter approach has two main advantages; 1) the complexity of the data obtained is much smaller, and 2) more sequences can be assigned to a specific organism or a group of related organisms. The latter facilitates some semi-quantitative analysis which is much more difficult when analyzing all sequences mixed from many organisms with largely varying genomesizes. Metagenomics focussing on rDNA genes has been used to study many different things, incl. e.g. the effect of the 2004 tsunami on microbial ecologies in marine, brackish, freshwater and terrestrial communities in Thailand [40]. Targeting the chloroplast trnL gene was successfully used to study airborne pollen and its relation with hay-fever [41].

A unique feature of metagenomics approaches is that one does not need to culture the organisms that one wants to study. The most impressive result of first studies that read all DNA sequences was that up to ten times more organisms were encountered than seen previously. One can now study organisms that no one is able to culture and/or that no one has ever seen. Based on DNA sequencing one gets an idea of the complexity and constitution of entire ecosystems [42,43]. Using RNA sequencing one gets an overview of "what's happening" [44]. The technology facilitates the study of the consequences of environmental changes as well as way to determine the cause of disturbances. Enormous progress has been made in a medical setting. Studying the human microbiome gave a range of surprising findings, incl. its enormous complexity containing 10× more cells than the human body and 1000× more genes. The microbiome of the human gut has been studied in great detail [45] including its relation with phenotypes like obesity [46]. Microbiomics is hot and it will undoubtedly bring new insights in the interplay between human health and the bugs living on and in within us.

5.6. Non-invasive prenatal testing

To obtain DNA from a fetus, prenatal diagnosis generally involves the costly and risky sampling of either chorionic villi or amniotic fluid. It is long known that DNA of the fetus can be found in maternal blood (cell free serum), yet it has low abundance and low quality and it is not easy to discriminate fetal from maternal DNA. These characteristics prevented wide-spread implementation of prenatal tests performed on maternal blood. However, the enormous power of NGS technology seems particularly attractive for non-invasive prenatal testing (NIPT). To detect trisomies, in particular trisomy-21 or Down's syndrome, a very simple but effective brute-force method was developed: sequence, map, and count. DNA isolated from maternal serum is sequenced, reads are mapped to the human genome and counted per chromosome. When 5–10 million reads are mapped, trisomies will reveal themselves by giving a significantly too high number of reads mapping to a particular chromosome [47]. A recent study showed that when genome sequencing of both parents, genome-wide maternal haplotyping and deep sequencing of maternal plasma DNA are combined even the genome sequence of an 18.5 weeks human fetus can be determined [48].

5.7. Disease gene identification

A combination of genome-wide association studies (GWAS) and specific targeting by sequence capture of the genomic regions detected is now used extensively trying to identify the variants that functionally link the DNA with the phenotype. Similarly, genome sequencing can be used as a tool to characterize genetic variation in a specific population, determine haplotype structure and use this knowledge to impute alleles and boost the outcome of GWAS analysis [GoNL consortium, 2014, in press].

One of the most impressive applications of NGS lies in the field of human genetics and disease gene identification. In the past larger families were an absolute requirement for a successful approach. Without being able to first map a disease gene to a specific position and then zooming in on the genes in that region, the human genome was simply too big and analysis too costly. Some successes were obtained using candidate disease gene approaches but generally these only worked when for a specific disease a new gene was discovered making similar genes or neighboring genes in a certain pathway obvious candidates. NGS studies were much more successful, even when only one or a few cases are available [49]. Parent–child trio analysis turned out to be very effective to reveal dominant de novo diseases [50], while recessive diseases can be revealed when several unrelated cases are available or when clearly damaging variants are present [51]. The latter successes could already be obtained with exome sequencing, i.e. a method to zoom in on the 1–2% protein coding sequences of the human genome only. Needless to say that, when cost drops further, full genome sequencing will be used to detect also deleterious variants that are not in the protein coding regions. Early steps towards the ultimate application, genome-based medicine/personalized medicine have been set by the UK and Saudi Arabia which last year both announced projects to sequence the genome of 100,000 individuals.

5.8. Human disease and health

Thus far studies have been mostly performed on the level of cell cultures, whole tissues or sorted cell populations. Although the yield per cell, 30%–70% of all RNA or DNA present, can still be improved, recent NGS developments have now made genome-wide single cell analysis feasible. Individual cells turn out to be quite different showing extensive genomic and transcriptomic heterogeneity in both normal development and disease [52]. This turns out to be especially true for cancer tissue being a complex mixture of many different cell populations each carrying a range of genomic rearrangements driving its unrestricted growth. Dissecting these using (very) deep sequencing of the cancer genome/transcriptome as a whole or from single cell analysis should give us a tool to identify the so-called driver mutations. These will be different in different tumors and instrumental to direct treatment and prescribe the best (set of) drugs to be used, personalized cancer treatment. Similarly NGS approaches will be used to study drug resistance, identify their mechanism and provide strategies to combat resistance [53]. Coordinated by the International Cancer Genome Consortium (ICGC) a large project is ongoing trying to resolve the genomic changes present in many forms of cancers by analyzing 50 cancer/normal tissue pairs [54]. In due time NGS developments will start to impact our daily life. While it will initially be used as a molecular microscope to diagnose disease, ultimately it will also be used to monitor our personal health. Our genome sequence will be read once, but e.g. blood-derived RNA analysis, completed with proteomics and metabolomics measurements, will be used on a regular basis to study the status of our body [55], the Whole-body-BIOscan.

5.9. Single molecule & long read sequencing

It should be noted that high accuracy sequences, i.e. sequences containing few read errors, are not essential for all applications. For specific studies, low accuracy (>85%) single-molecule long-read sequences can be sufficient to make a significant difference. De novo genome assembly can be improved considerably when long kilobase-sized reads are available that span gaps from short-read paired-end sequencing. The single-molecule technologies are amplification free and thereby not hampered by PCR-based artifacts like uneven amplification and GC-bias. Consequently, they give a much more uniform coverage and span GC-rich regions. In addition they may span repeat structures or duplications that cannot be resolved using short read sequencing. English et al. have used PacBio data for upgrading existing draft genome assemblies derived from Illumina sequencing data by looking for reads that extend into or cover gaps in the assembly [56]. A paper by Loomis et al. showed that the system has no problems sequencing long regions of 100% GC content of CGG trinucleotide repeat expansions [57].

Amplification-free methods facilitate the analysis of DNA modifications, deriving from either cellular processes like methylation [58] and/or from damage (irradiation, chemicals, etc) [59]. DNA modifications present on the template molecule affect the DNA polymerase activity to a certain extent, i.e., the time needed for incorporation of a nucleotide at a particular site. The identity of modifications can be inferred by analysis of the kinetic state of the polymerase. This works well for modifications that have a large effect on the polymerase activity, e.g., N6-methyladenine (m6A) and N4-methylcytosine (m4C) [60,61]. The 5-methylcytosine (m5C) modification has a weaker signal but given enough coverage can still be inferred from the data [61]. However, conversion of 5-methylcytosine to 5-carboxylcytosine through the Ten-eleven Translocation Gene Protein 1 (Tet1) results in a greater disturbance in the signal, thus making it easier to detect even at lower coverage [62].

Other applications of long-read sequencing are RNA structure analysis [63] and studies to unravel complex repeat structures (e.g. segmental duplications in the human genome) and large segments of repetitive DNA.

6. Online supplement

6.1. Solid sequencing

The Solid systems read DNA by an intricate sequencing-by-ligation scheme [64]. After positioning the templated beads on the flow-cell, the first step is hybridization of a primer complementary to the complete adapter followed by hybridization of octamer probes. The first two nucleotides of these probes represent 16 dinucleotide combinations, bases 3–5 are degenerate and bases 6–8, also degenerate, contain a fluorescent label for identification of bases 1 and 2. Four different dinucleotide combinations are labeled with the same fluorescent tag. Use of 4 different fluorescent groups allows for labeling all 16 dinucleotide tags. When a probe anneals adjacent to the adapter primer strands are ligated and information on the first two bases is recorded from the fluorescent signal at bases 6–8. These last three bases are removed and new octomer probes are allowed to hybridize and ligate, providing information on bases 6 and 7, i.e., position +5 relative to the previous cycle, of the template. Depending on the required read length, 5–7 of such cycles are performed after which the entire synthesized strand is removed from the template. A new primer with a −1 nt shift in end position is hybridized and octomer probes are allowed to hybridize and ligate in 5–7 cycles as described above, but now probing nucleotides 0–1, 5–6, and 10–11 etc. of the template. In total 5 of these re-priming cycles are performed. The template sequence is subsequently decoded from the color labels from the two ligation events per base [65].

Similar to Illumina sequencing, the Solid system is able to generate paired-end reads. However, due to limitations of sequencing by ligation scheme the maximal read length is limited to 75 bases for read 1 and 35 bases for read 2. Although the total number of reads generated per Solid run is comparable to HiSeq, the total output in Gb per run is only half of that for HiSeq due to the shorter read length. The main advantage of this

scheme is that each base is interrogated by two octomer ligations, which results in a significant increase in read accuracy which may be as low as 0.01%. By adding yet another re-priming cycle, the accuracy can theoretically be improved even more.

6.2. Roche 454

The Roche company has produced two Pyrosequencing [66] platforms, i.e., the FLX + and the bench top GS Junior system, on the same technology. Similar to Ion torrent sequencing, templated beads are deposited in a picotiter plate to separate the individual sequence reactions and T–A–C–G nucleotide flows are applied to the plate. The main difference is that the sequencing reaction is monitored by detection light generated by luciferase-mediated conversion of luciferin to oxyluciferin upon primer extension. For a long time, the main niche for these systems was the long read lengths of 500–800 bp. However, increased read lengths for competitor platforms have made pyrosequencing less cost efficient. The Roche systems were announced to be phased out in mid-2016.

References

[1] J. Dabney, M. Meyer, Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries, Biotechniques 52 (2012) 87–94.
[2] M. Hafner, N. Renwick, M. Brown, A. Mihailović, D. Holoch, C. Lin, J.T. Pena, J.D. Nusbaum, P. Morozov, J. Ludwig, et al., RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries, RNA 17 (2011) 1697–1712.
[3] R.D. Mitra, G.M. Church, In situ localized amplification and contact replication of many individual DNA molecules, Nucleic Acids Res. 27 (1999) e34–e39.
[4] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.
[5] M. Nakano, J. Komatsu, S.-i. Matsuura, K. Takashima, S. Katsura, A. Mizuno, Single-molecule pcr using water-in-oil emulsion, J. Biotechnol. 102 (2003) 117–124.
[6] D. Dressman, H. Yan, G. Traverso, K.W. Kinzler, B. Vogelstein, Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations, Proc. Natl. Acad. Sci. 100 (2003) 8817–8822.
[7] D. Golan, P. Medvedev, Using state machines to model the ion torrent sequencing process and to improve read error rates, Bioinformatics 29 (2013) i344–i351.
[8] B. Merriman, IonTorrentR&D-team, J.M. Rothberg, Progress in ion torrent semiconductor chip based sequencing, Electrophoresis 33 (2012) 3397–3417.
[9] Z. Ma, R.W. Lee, B. Li, P. Kenney, Y. Wang, J. Erikson, S. Goyal, K. Lao, Isothermal amplification method for next-generation sequencing, Proc. Natl. Acad. Sci. 110 (35) (2013) 14320–14323.
[10] P. Coupland, T. Chandra, M. Quail, W. Reik, H. Swerdlow, Direct sequencing of small genomes on the pacific biosciences RS without library preparation, Biotechniques 53 (2012) 365–372.
[11] M.J. Levene, J. Korlach, S.W. Turner, M. Foquet, H.G. Craighead, W.W. Webb, Zero-mode waveguides for single-molecule analysis at high concentrations, Science 299 (2003) 682–686.
[12] J. Korlach, P.J. Marks, R.L. Cicero, J.J. Gray, D.L. Murphy, D.B. Roitman, T.T. Pham, G.A. Otto, M. Foquet, S.W. Turner, Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures, Proc. Natl. Acad. Sci. 105 (2008) 1176–1181.
[13] P.M. Lundquist, C.F. Zhong, P. Zhao, A.B. Tomaney, P.S. Peluso, J. Dixon, B. Bettman, Y. Lacroix, D.P. Kwo, E. McCullough, et al., Parallel confocal detection of single molecules in real time, Opt. Lett. 33 (2008) 1026–1028.
[14] M. de Vega, J.M. Lazaro, M. Salas, L. Blanco, Primer-terminus stabilization at the 3′–5′ exonuclease active site of phi29 DNA polymerase. Involvement of two amino acid residues highly conserved in proofreading DNA polymerases, EMBO J. 15 (1996) 1182.
[15] J. Esteban, M. Salas, L. Blanco, Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization, J. Biol. Chem. 268 (1993) 2719–2726.
[16] L. Blanco, A. Bernad, J.M. Lázaro, G. Martin, C. Garmendia, M. Salas, Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication, J. Biol. Chem. 264 (1989) 8935–8940.
[17] C.-S. Chin, D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, et al., Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Nat. Methods 10 (6) (2013) 563–569.
[18] N. Ashkenasy, J. Snchez-Quesada, H. Bayley, M.R. Ghadiri, Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores, Angew. Chem. Int. Ed. 44 (2005) 1401–1404.
[19] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, H. Bayley, Continuous base identification for single-molecule nanopore DNA sequencing, Nat. Nanotechnol. 4 (2009) 265–270.
[20] J. Li, M. Gershow, D. Stein, E. Brandin, J.A. Golovchenko, DNA molecules and configurations in a solid-state nanopore microscope, Nat. Mater. 2 (2003) 611–615.
[21] D. Fologea, M. Gershow, B. Ledden, D.S. McNabb, J.A. Golovchenko, J. Li, Detecting single stranded DNA with a solid state nanopore, Nano Lett. 5 (2005) 1905–1909.
[22] P. A.'t Hoen, Y. Ariyurek, H.H. Thygesen, E. Vreugdenhil, R.H. Vossen, R.X. de Menezes, J.M. Boer, G.-J.B. van Ommen, J.T. den Dunnen, Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms, Nucleic Acids Res. 36 (2008) e141.
[23] T. Lappalainen, M. Sammeth, M.R. Friedlander, P.A.C. 't Hoen, J. Monlong, M.A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P.G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D.G. MacArthur, M. Lek, E. Lizano, H.P.J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S.B. Montgomery, P. Donnelly, M.I. McCarthy, P. Flicek, T.M. Strom, T.G. Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S.E. Antonarakis, R. Hasler, A.-C. Syvanen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I.G. Gut, X. Estivill, E.T. Dermitzakis, Transcriptome and genome sequencing uncovers functional variation in humans, Nature 501 (2013) 506–511.
[24] H.J. Westra, M.J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M.W. Christiansen, B.P. Fairfax, K. Schramm, J.E. Powell, A. Zhernakova, D.V. Zhernakova, J.H. Veldink, L.H. Van den Berg, J. Karjalainen, S. Withoff, A.G. Uitterlinden, A. Hofman, F. Rivadeneira, P.A.C. 't Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A.B. Singleton, D.G. Hernandez, M.A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Volker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S.A. Gharib, D.A. Enquobahrie, T. Lumley, G.W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R.C. Jansen, P.M. Visscher, J.C. Knight, B.M. Psaty, S. Ripatti, A. Teumer, T.M. Frayling, A. Metspalu, J.B.J. van Meurs, L. Franke, Systematic identification of trans eQTLs as putative drivers of known disease associations, Nat. Genet. 45 (2013) 1238–1243.
[25] E. Valen, G. Pascarella, A. Chalk, N. Maeda, M. Kojima, C. Kawazu, M. Murata, H. Nishiyori, D. Lazarevic, D. Motti, et al., Genome-wide detection and analysis of hippocampus core promoters using deepcage, Genome Res. 19 (2009) 255–265.
[26] E. de Klerk, A. Venema, S.Y. Anvar, J.J. Goeman, O. Hu, C. Trollet, G. Dickson, J.T. den Dunnen, S.M. van der Maarel, V. Raz, et al., Poly (a) binding protein nuclear 1 levels affect alternative polyadenylation, Nucleic Acids Res. 40 (2012) 9089–9101.
[27] E. Berezikov, F. Thuemmler, L.W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, R.H.A. Plasterk, Diversity of microRNAs in human and chimpanzee brain, Nat. Genet. 38 (2006) 1375–1377.
[28] E.N.M. Nolte-'t Hoen, H.P.J. Buermans, M. Waasdorp, W. Stoorvogel, M.H.M. Wauben, P.A.C. t Hoen, Deep sequencing of RNA from immune cell-derived vesicles uncovers the selective incorporation of small non-coding RNA biotypes with potential regulatory functions, Nucleic Acids Res. 40 (2012) 9272–9285.
[29] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F.O. Bagger, M. Jorgensen, P.R. Andersen, N. Bertin, O. Rackham, A.M. Burroughs, J.K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C.J. Mungall, T.F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C.O. Daub, P. Heutink, D.A. Hume, T.H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, T.F. Consortium, A.R.R. Forrest, P. Carninci, M. Rehli, A. Sandelin, An atlas of active enhancers across human cell types and tissues, Nature 507 (2014) 455–461.
[30] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, J.S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling, Science 324 (2009) 218–223.
[31] Y. Blat, N. Kleckner, Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region, Cell 98 (1999) 249–259.
[32] E.P. Consortium, et al., The encode (encyclopedia of DNA elements) project, Science 306 (2004) 636–640.
[33] P.W. Laird, Principles and challenges of genome-wide DNA methylation analysis, Nat. Rev. Genet. 11 (2010) 191–203.
[34] B.R. Herb, F. Wolschin, K.D. Hansen, M.J. Aryee, B. Langmead, R. Irizarry, G.V. Amdam, A.P. Feinberg, Reversible switching between epigenetic states in honeybee behavioral subcastes, Nat. Neurosci. 15 (2012) 1371–1373.
[35] R. Li, W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, et al., The sequence and de novo assembly of the giant panda genome, Nature 463 (2009) 311–317.
[36] N. Rohland, D. Reich, S. Mallick, M. Meyer, R.E. Green, N.J. Georgiadis, A.L. Roca, M. Hofreiter, Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants, PLoS Biol. 8 (2010) e1000564.
[37] D. Reich, R.E. Green, M. Kircher, J. Krause, N. Patterson, E.Y. Durand, B. Viola, A.W. Briggs, U. Stenzel, P.L. Johnson, et al., Genetic history of an archaic hominin group from denisova cave in Siberia, Nature 468 (2010) 1053–1060.
[38] D.L. Cox-Foster, S. Conlan, E.C. Holmes, G. Palacios, J.D. Evans, N.A. Moran, P.-L. Quan, T. Briese, M. Hornig, D.M. Geiser, et al., A metagenomic survey of microbes in honey bee colony collapse disorder, Science 318 (2007) 283–287.
[39] K.I. Bos, V.J. Schuenemann, G.B. Golding, H.A. Burbano, N. Waglechner, B.K. Coombes, J.B. McPhee, S.N. DeWitte, M. Meyer, S. Schmedes, et al., A draft genome of *Yersinia pestis* from victims of the black death, Nature 478 (2011) 506–510.
[40] N. Somboonna, A. Wilantho, K. Jankaew, A. Assawamakin, D. Sangsrakru, S. Tangphatsornruang, S. Tongsima, Microbial ecology of Thailand tsunami and non-tsunami affected terrestrials, PLoS One 9 (2014) e94236.
[41] K. Kraaijeveld, L.A. de Weger, M. Ventayol GarcÃ-a, H. Buermans, J. Frank, P.S. Hiemstra, J.T. den Dunnen, Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing, Mol. Ecol. Resour. (2014), http://dx.doi.org/10.1111/1755-0998.12288 (http://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12288/abstract).

[42] J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, et al., Environmental genome shotgun sequencing of the Sargasso sea, Science 304 (2004) 66–74.
[43] D. Ercolini, F. De Filippis, A. La Storia, M. Iacono, Remake by high-throughput sequencing of the microbiota involved in the production of water buffalo mozzarella cheese, Appl. Environ. Microbiol. 78 (2012) 8142–8145.
[44] Z. Wang, M. Gerstein, M. Snyder, RNA-seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[45] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, et al., Enterotypes of the human gut microbiome, Nature 473 (2011) 174–180.
[46] P.J. Turnbaugh, F. Bäckhed, L. Fulton, J.I. Gordon, Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome, Cell Host Microbe 3 (2008) 213–223.
[47] M.E. Norton, H. Brar, J. Weiss, A. Karimi, L.C. Laurent, A.B. Caughey, M.H. Rodriguez, J. W. III, M.E. Mitchell, C.D. Adair, H. Lee, B. Jacobsson, M.W. Tomlinson, D. Oepkes, D. Hollemon, A.B. Sparks, A. Oliphant, K. Song, Non-invasive chromosomal evaluation (nice) study: results of a multicenter prospective cohort study for detection of fetal trisomy 21 and trisomy 18, Am. J. Obstet. Gynecol. 207 (2012) 137.e1–137.e8.
[48] J.O. Kitzman, M.W. Snyder, M. Ventura, A.P. Lewis, R. Qiu, L.E. Simmons, H.S. Gammill, C.E. Rubens, D.A. Santillan, J.C. Murray, H.K. Tabor, M.J. Bamshad, E.E. Eichler, J. Shendure, Noninvasive whole-genome sequencing of a human fetus, Sci. Transl. Med. 4 (2012) 137ra76.
[49] A. Hoischen, B.W.M. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, P. de Vries, R. de Reuver, N. Wieskamp, G. Mortier, K. Devriendt, M.Z. Amorim, N. Revencu, A. Kidd, M. Barbosa, A. Turner, J. Smith, C. Oley, A. Henderson, I.M. Hayes, E.M. Thompson, H.G. Brunner, B.B.A. de Vries, J.A. Veltman, De novo mutations of SETBP1 cause Schinzel–Giedion syndrome, Nat. Genet. 42 (2010) 483–485.
[50] L.E.L.M. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, B.W.M. van Bon, A. Hoischen, B.B.A. de Vries, H.G. Brunner, J.A. Veltman, A de novo paradigm for mental retardation, Nat. Genet. 42 (2010) 1109–1112.
[51] S.B. Ng, K.J. Buckingham, C. Lee, A.W. Bigham, H.K. Tabor, K.M. Dent, C.D. Huff, P.T. Shannon, E.W. Jabs, D.A. Nickerson, J. Shendure, M.J. Bamshad, Exome sequencing identifies the cause of a Mendelian disorder, Nat. Genet. 42 (2010) 30–35.
[52] R. Bernards, Finding effective cancer therapies through loss of function genetic screens, Curr. Opin. Genet. Dev. 24 (2014) 23–29.
[53] I.C. Macaulay, T. Voet, Single cell genomics: advances and future perspectives, PLoS Genet. 10 (2014) e1004126.
[54] J. Zhang, J. Baran, A. Cros, J.M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, A. Kasprzyk, International cancer genome consortium data portal: a one-stop shop for cancer genomics data, Database 2011 (2011) bar026 http://dx.doi.org/10.1093/database/bar026.
[55] R. Chen, G. Mias, J. Li-Pook-Than, L. Jiang, H. Lam, R. Chen, E. Miriami, K. Karczewski, M. Hariharan, F. Dewey, Y. Cheng, M. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. Blasco, P. Greenberg, P. Snyder, T. Klein, R. Altman, A.J. Butte, E. Ashley, M. Gerstein, K. Nadeau, H. Tang, M. Snyder, Personal omics profiling reveals dynamic molecular and medical phenotypes, Cell 148 (2012) 1293–1307.
[56] A.C. English, S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D.M. Muzny, J.G. Reid, K.C. Worley, et al., Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology, PLoS One 7 (2012) e47768.
[57] E.W. Loomis, J.S. Eid, P. Peluso, J. Yin, L. Hickey, D. Rank, S. McCalmon, R.J. Hagerman, F. Tassone, P.J. Hagerman, Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene, Genome Res. 23 (2013) 121–128.
[58] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, S.W. Turner, Direct detection of DNA methylation during single-molecule, real-time sequencing, Nat. Methods 7 (2010) 461–465.
[59] T.A. Clark, K.E. Spittle, S.W. Turner, J. Korlach, et al., Direct detection and sequencing of damaged DNA bases, Genome Biol. 2 (2011).
[60] I.A. Murray, T.A. Clark, R.D. Morgan, M. Boitano, B.P. Anton, K. Luong, A. Fomenkov, S.W. Turner, J. Korlach, R.J. Roberts, The methylomes of six bacteria, Nucleic Acids Res. 40 (2012) 11450–11462.
[61] G. Fang, D. Munera, D.I. Friedman, A. Mandlik, M.C. Chao, O. Banerjee, Z. Feng, B. Losic, M.C. Mahajan, O.J. Jabado, et al., Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing, Nat. Biotechnol. 30 (12) (2012) 1232–1239.
[62] T.A. Clark, X. Lu, K. Luong, Q. Dai, M. Boitano, S.W. Turner, C. He, J. Korlach, Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation, BMC Biol. 11 (2013) 4.
[63] D. Sharon, H. Tilgner, F. Grubert, M. Snyder, A single-molecule long-read survey of the human transcriptome, Natl. Biotechnol. Adv. Online Publ. 31 (11) (2013) 1009–1014.
[64] J. Shendure, G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, G.M. Church, Accurate multiplex polony sequencing of an evolved bacterial genome, Science 309 (2005) 1728–1732.
[65] M.L. Metzker, Sequencing technologies-the next generation, Nat. Rev. Genet. 11 (2009) 31–46.
[66] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.-J. Chen, Z. Chen, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.