Accepted Manuscript

A glimpse into past, present, and future DNA sequencing

Marcos Morey, Ana Fernández-Marmiesse, Daisy Castiñeiras, José M. Fraga, María L. Couce, José A. Cocho

PII: S1096-7192(13)00154-6
DOI: doi: 10.1016/j.ymgme.2013.04.024
Reference: YMGME 5544

To appear in: Molecular Genetics and Metabolism

Received date: 27 February 2013
Revised date: 30 April 2013
Accepted date: 30 April 2013

Please cite this article as: Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J.M., Couce, M.L. & Cocho, J.A., A glimpse into past, present, and future DNA sequencing, Molecular Genetics and Metabolism (2013), doi: 10.1016/j.ymgme.2013.04.024

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
A glimpse into past, present, and future DNA sequencing
Marcos Morey, Ana Fernández-Marmiesse, Daisy Castiñeiras, José M. Fraga, María L. Couce, José A. Cocho.

Marcos Morey
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Hospital Clínico Universitario de Santiago. A Choupana s/n, 15706 Santiago de Compostela. Spain.marcosmorey84@gmail.com

Ana Fernandez-Marmiesse
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Hospital Clínico Universitario de Santiago A Choupana s/n, 15706 Santiago de Compostela. Spain.amarmiesse@gmail.com

Daisy Castiñeiras
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Hospital Clínico Universitario de Santiago A Choupana s/n, 15706 Santiago de Compostela. Spain. Daisy.Castineiras.Ramos@sergas.es

José M. Fraga
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Departamento de Pediatría. Hospital Clínico Universitario de Santiago. Universidad de Santiago de Compostela. A Choupana s/n, 15706 Santiago de Compostela. Spain. pdneonat@usc.es

María L. Couce: corresponding author
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Departamento de Pediatría. Hospital Clínico Universitario de Santiago. Universidad de Santiago de Compostela. A Choupana s/n, 15706 Santiago de Compostela. Spain e-mail: maria.luz.couce.pico@sergas.es.

José A. Cocho
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Hospital Clínico Universitario de Santiago A Choupana s/n, 15706 Santiago de Compostela. Spain. cochoja@yahoo.es

**Corresponding author:** María L. Couce
Unidad de Diagnóstico y Tratamiento de Enfermedades Metabólicas Congénitas. Departamento de Pediatría. Hospital Clínico Universitario de Santiago. Universidad de Santiago de Compostela. A Choupana s/n, 15706 Santiago de Compostela. Spain e-mail: maria.luz.couce.pico@sergas.es. Telephone. 0034981950162
A Glimpse into Past, Present, and Future DNA Sequencing

Abstract
Current advances in DNA sequencing technologies are dropping down sequencing cost while increasing throughput at a pace never shown before. Past-decade great milestones, as the establishment of a reference human genome (amongst others) and large-scale human genetic variation study in the 1000 Genome project are, in conjunction with the use of these techniques, triggering advances in many areas of basic and applied science. These tools, stored in and combined with the vast amount of information present in biological online databases are, with the use of automated interpretation and analysis tools, allowing the fulfillment of increasingly ambitious studies in many areas and also are democratizing the access to information, interpretation and technologies, being the first opportunity for researchers to assess the influence of genetics in complex events as multifactorial diseases, evolutionary studies, metagenomics, transcriptomics, etc. In this review, we present the current state of the art of these technologies, focusing on second generation sequencing, from sample and library preparation to sequencing chemistries and bioinformatic software available for final data analysis and visualization, with its possible applications. We also make an overview of first and third generation, due to its historical importance and for being the upcoming future tools for genetic analysis, respectively.

Keywords
Next generation sequencing; DNA sequencing; genome sequencing; high throughput genomics; third generation sequencing; sequencing technologies.

As we move into an era of personalized medicine and complex genomic databases, the demand for development of new and existing sequencing technologies is constant. Although we can not yet sequence an individual genome for $1000, novel approaches are reducing the cost per base and increasing throughput on a daily basis[1]. The establishment of the human reference genome in 2001[2, 3] and the achievements of large sequencing projects such as the 1000 Genome Project[4] are catalysing advances in human genetics. A broad range of areas of basic and applied research are benefitting from such advances, notably genetic diagnostics[5], biotechnology[6], microbiological studies[7], forensic biology[8], and systematics and taxonomy[9, 10]. Advances in sequencing methodologies, moreover, are changing the ways in which scientists analyse and understand genomes, while the results that they yield are disseminated increasingly widely through science news magazines[11].

In the medical sciences, advances in our knowledge of the genetic basis of pathologies have changed the way in which such entities are understood. Thus, disease has gone from being an individual-specific to a familial phenomenon, in which genetic alterations (mutations) can be genealogically traced [12]. Such progress has led to the development of numerous applications for disease prevention and diagnosis, prenatal screening[13], and genetic counseling.
Historically, distinct generations of sequencing chemistries can be distinguished. Classical methods began in the 1970s[14, 15], when the first enzymatic methods became widely used for most applications[16]. Due to their inherently limited throughput, however, large studies are not feasible that methodologies. New technologies, known as Next-Generation Sequencing (NGS) technologies, or Second Generation Sequencing, have recently been developed[17]. Using these, it is possible to analyze complete genomes[18], exomes[19], and large gene panels[20], complete RNA transcriptomes[21], and also to carry out a range of other techniques, including chromatin immunoprecipitation[22]. In medical practice, complex or rare diseases that previously were assessed in terms of one or a few putative causal genes can now be more thoroughly analyzed. Using modern technologies it is thus possible to carry out more rigorous research with unambiguous outcomes and to diagnose genetic-based diseases[23, 24].

In this review, we describe the state of the art of NGS technologies and their potential applications. We also briefly summarize the current state of Third Generation Sequencing technologies, still very much in their infancy[25].

**Classical sequencing: First Generation**

**Sanger Sequencing**

In the 70s, Sanger and Coulson developed what is known as enzymatic DNA sequencing, or Sanger Sequencing which utilises DNA polymerase [16], in contrast to previous non-enzymatic approaches[15]. In 1975 they published their sequencing method, known as ‘plus and minus’ and the sequence of bacteriophage φX174 [14]. Two years later the same authors presented the ‘chain termination method’, a more efficient and easier variation on the ’plus and minus’ method[16] which utilised radioactively or fluorescently[26] labeled deoxyribonucleotides acting as chain terminators[27]. The template DNA to be sequenced is obtained by in vivo (Figure 01. A) or in vitro cloning (Figure 01. B) and added to four parallel reactions. Each reaction contains four unmarked deoxyxynucleotides and one deoxyxynucleotide marked by its lack of 3’OH, which impedes chain elongation. The template elongates by incorporation of deoxyxynucleotides in random positions: this process generates amplicons of any size which can then be spatially separated in parallel acrylamide gels to obtain a sequence of up to a few hundred bases[16].

This technique has been since been optimised as follows: 1) using a nucleotide-specific fluorescent dye, which allows the chain-termination reaction for the four deoxyxynucleotides to be carried out in a single reaction (Figure 01. C) [26]; 2) replacement of conventional acrylamide gels with polyacrylamide gels in capillary electrophoresis, to shorten run times[28]; 3) automatic laser fluorescence detection (Figure 01. D)[29].

Even today, multiplexing and miniaturization are being used to reduce reagent volumes and consumable costs while increasing throughput. Thus, in microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments and separation by electrophoresis is done on a single slide of approximately 10cm diameter. While this reduces reagent usage as well as cost[30], further work is needed to make this variation of Sanger Sequencing commercially viable[31].

**Maxam and Gilbert Sequencing**

In the mid 70s, Maxam and Gilbert developed a non-enzymatic DNA sequencing method. Samples to be sequenced require less complex preparation than
in Sanger Sequencing, but reads are shorter (maximum 100bp). The process itself is technically complex and involves the use of hazardous chemicals.

The method consists of selectively fragmenting the region to be sequenced, for each of the four nitrogenous bases or combinations thereof, and loading the resulting fragments onto a polyacrylamide gel. Depending on the sizes of the fragments and how they are cut, the radioactively labeled fragments can be read to obtain the sequence in a similar way to Sanger sequencing [15].

Limitations of classic sequencing techniques

The main limitation of both these classical sequencing techniques is their low throughput, due to template preparation and in the case of Sanger Sequencing also to carrying out the enzymatic reaction. Each run in Sanger Sequencing can sequence up to 1000bp, and with an automated sequencer 384 sequences can be run in parallel with a throughput of 80-100kb per hour. Due its singleplex nature, Sanger Sequencing is not a hardly scalable process. In 1985, reading a single base cost $10, while in 2005, the various improvements reading 10,000 bases cost the same. However, large projects such as the Human Genome Project still required vast amounts of time and resources[2]. Another limitation of First Generation Sequencing is that variants present at low frequency, such as mosaics, are difficult to detect due to high background levels. Finally, compared with modern technologies, the cost per base is still high [32].

The above limitations are important to consider when carrying out studies using these older technologies. However, due to the development of new approaches to DNA sequencing in which millions or even billions of reads can be carried out in parallel, the scenario is changing, to allow a large number of sequences to be obtained in a short time.

Next Generation Sequencing: Second Generation

Introduction

The first NGS equipment became available in 2004. The chance to sequence an entire genome at an unprecedented speed and for a reasonable cost was used to address previously intangible questions, for example differential susceptibility to disease among individuals[33], the influence of the microbiome on health [34], understanding cancer genomes [35], the roles of rare variants in common diseases [36]. These new avenues are also revealing previously unknown genomic variation [37, 38].

Indeed, NGS DNA sequencing is somewhat of a tipping point in research. Previously, research questions were ahead of what technology could address. But now, technology can not only address previously formulated biological questions, but is rapidly giving rise to new ones.

All NGS chemistries have in common the cyclic parallel reading of clonally amplified and spatially separated amplicons[39]. There is a significant conceptual change with these new approaches: in classical sequencing, a single amplicon from a single sample is amplified and a unique sequence is obtained. If 100% of the sample is amplified, 100% can be sequenced each time. NGS chemistry, on the other hand, is a massive representation of many random events. While it is impossible to drive them all in one direction, the enormous size of the process means that all sequences are likely to be represented. These processes usually follow a Poisson distribution.
There are key factors in the definition and evaluation of any Next-Generation technology platform: 1) read length, 2) throughput, 3) read accuracy (measurable using the Phred Score[40]), 4) read depth (number of times each base is sequenced in independent events), and 5) cost per base. These methodologies have brought down the costs of sequencing per base in recent years, from $0.01 in 2004 to about $0.0001 in 2006, and these continue to fall as capacity rises[1, 32] (Figure 02[41]). Thus, the next bottleneck to consider will be the management and interpretation of the vast amount of data generated[17].

The main weakness of NGS is that the read length is not as long as with the previous techniques, due mainly to the progressive decline in efficiency of the sequencing chemistry during the run. This also results in an asynchronous read elongation for any given amplicon clone (dephasing or prephasing) and a correspondingly ambiguous fluorescent signal[42]. This asynchrony increases with the number of sequencing cycles, finally reaching a point at which the data of those cycles are unusable. Short reads are a problem because they are difficult to align with the reference genome. Another weakness of NGS is the use of PCR which can itself potentially introduce bias[43].

NGS equipment provides a streamlined sample pipeline in comparison to previous technologies. In fact, one of the key advances of NGS sample preparation is its capacity to perform highly multiplexed reactions. Classical DNA sequencing, due to its ‘one sample-one PCR-one sequence’  nature only allows uniplex reactions.

The objective of NGS sample preparation is clonal amplification of single strands of the initial DNA. Thus, relatively significant amounts (µg) of initially high molecular weight DNA, with high purity levels (260nm/280nm=1.8, 260nm/230nm=2), are required. However, these levels depend on the application, and newer protocols tend to be less strict, enabling the use of DNA from a wide range of sources (paraffin-wax, tissue, frozen blood, stained paper, etc).

This review focuses mainly on DNA sequencing, with the techniques described below being applied mainly to those ends, rather than to the wide range of other sequencing applications. The different types of DNA libraries used in NGS will be described (Figure 03). These are usually obtained by annealing specific DNA adaptors/bases to fragmented and size-selected DNA, although the precise mechanism of preparation depends on sequencing chemistry and the size of the region of interest.

Regardless of technology, there are several common steps in DNA sequencing. First, the DNA sample is usually fragmented to a target size (150 to many hundred base pairs), depending on the platform read length and chemistry used. This fragmentation is usually carried out by mechanical methods, i.e. sonication or nebulization. The resulting fragment sizes follow a Poisson distribution. Currently, this step is the first bottleneck in processing a large number of samples. Methods of enzymatic sample fragmentation are currently being commercialized, offering the possibility of easily parallelizing the process[44]. DNA fragment-size distribution can be optionally refined via gel-based size selection[45] or SPRI bead-size selection[46], to improve library quality[47]. Universal DNA adaptors or bases can be ligase-bound at both ends of each DNA fragment, to allow PCR amplification to be performed either with a single pair of primers or to attach library fragments to surfaces by using adaptor-complementary oligos. Unlike in traditional sequencing, bacterial cloning is therefore no longer required. An additional variant is the addition of poly-A tails to one end of the fragment; this is used mainly in single molecule sequencing (SMS) technologies, which will be described below[48]. To minimise the introduction of PCR-based biases and maintain coverage uniformity, during sample preparation,
PCRs are performed with as few cycles as possible [49]. Thus, each technology prepares DNA fragments from the initial sample as a so-called fragment library, on which to perform single-strand clonal amplification and sequencing reactions.

Currently, different types of DNA fragment libraries can be distinguished: 1) a fragment Library; this consists of variable-size DNA fragments with universal adaptors attached at both ends (Figure 03 A). From each DNA fragment in the library a single read of variable length is obtained (Figure 03 A1). Oversampling at both ends instead of only at one, to generate paired-end reads can be considered an improvement on the Fragment library (Figure 03 A2). Fragment libraries and Paired-end libraries are similarly built, but Paired-end can be used to obtain the information necessary to improve the detection of small indels. Repetitive regions are difficult to process in either case[39]. 2) Barcoded Library; the addition of a specific DNA sequence (‘barcode’) to each fragment allows sample multiplexing (Figure 03 B) [50]. These can be added during library construction, being attached as adaptors to the DNA fragments by ligation or polymerization during an amplification process with barcoded primers. Barcodes are thus sample-specific sequences that allow an increase in the number of samples per run by allowing the platform to distinguish them spatially. The platform throughput is divided between the total number of samples[51, 52]. As with Fragment Libraries, a single tag (Figure 03 B1) or a pair-end tag can be obtained (Figure 03. B2). Barcodes also allow sample pooling during library preparation, saving time and reagents. 3) Mate-Pair Library; this consists of two DNA fragments, commonly separated by an internal adaptor and flanked by universal adaptors (Figure 03. C). Due to its construction mechanism both fragments are ‘mates’, being derived from the ends of the same initial fragment, and the distance between them is known. This method therefore allows colocalization in the reference sequence and comparison of distances between readings, and is suited to detection of large rearrangements and analysis of repetitive regions[53].

The utility of each type library depends on the experiment. For targeted resequencing of specific genome regions a Fragment Library (with or without pair-end sequencing) is the adequate. However, for the sequencing of complex or repetitive regions, detection of large rearrangements, or de novo genome assembly, Mate-Pair Libraries are preferred because of their capacity for sequence colocalization.

While the major advantage of these new technologies is their high throughput, for most applications it is neither necessary nor cost-effective to sequence the entire genome. Targeted enrichment, in which selected genomic regions are sequenced, are therefore often used for sequencing gene panels on demand. As less data is required per sample, Barcoded Libraries are used to process many samples per run.

Several approaches to target enrichment have been developed, which can differ with respect to several parameters: 1) sensitivity, or percentage of bases represented; 2) specificity, or percentage of the total sequence output that maps to the target regions; 3) uniformity, i.e. variability in sequence coverage across target regions; 4) reproducibility, or how well results correlate between replicates; 5) cost; 6) ease of use; and 7) the amount of DNA required[47, 54].

**Targeted Enrichment Strategies**

**Enrichment by PCR**

Due to its ability to amplify a single DNA molecule with high specificity, PCR has been the most widely used technique for processing samples for sequencing; indeed Sanger sequencing relies on the reading of single amplicons[55]. For its application to NGS it is necessary to increase PCR throughput.
Long range uniplex PCR (up to 20kb) is commercialized by several companies (e.g. Invitrogen Sequlprep[56], Qiagen SeqTarget[57]). These have the inconvenience that, before fragment library preparation, each reaction has to be cleaned, carefully normalized and pooled (Figure 04 A). Depending on the application, either short (100-200bp) or long amplicons can be prepared (Figure 04 B); the latter are required for example in rapid screening of mutational hot spots.

Nano-scale uniplex PCR is commercialized by Fluidigm[58] and Raindance[59]. The Fluidigm platform comprises thousands of singleplex PCRs in parallel on a microfluidic PCR chip. Raindance, through its platform Rainstorm, is based on emulsion PCR[60] and uses microdroplet PCR. Each droplet supports an independent PCR, and a droplet population can thus carry out thousands of simultaneous reactions each with specific primer-pairs. After thermal cycling, the emulsion is broken down and the products are subjected to library construction [61].

An increasingly popular alternative to the options above is multiplex PCR, in which, the number of reactions is kept as low as possible for a high throughput. Multiplex PCR requires a careful primer design in order to allow several dozen primers per reaction without any primer interactions or non-specific amplifications.

Compared with other enrichment techniques, such as those which are probe-hybridization based, targeted enrichment via PCR does not require such a high quality or quantity of DNA for sequencing regions of up to several hundred kilobases. It provides a very high enrichment ratio with few off-target reads. In the case of small amplicons it is also possible to bypass the initial step of sample DNA fragmentation[54]. If the regions to enrich are larger, the amount of DNA required grows linearly with the number of PCRs required.

The main weakness of target enrichment by PCR is that coverage uniformity may not be as good as with other techniques, being the correct normalization of PCR products crucial to improve it [47]. In the beginning of this methodology, it was difficult to perform enrichment by this approach if the region of interest involved many dispersed amplicons, due to issues with PCR-multiplexing, so it was more suitable to enrich kb-range contiguous regions due to the possibility to overlap the amplicons [62]. Currently, as the multiplexing capacity is being increased by the improvement and automation of primer-designing, enrichment by PCR of numerous dispersed genomic regions is possible, being it commercialized by different companies, e.g: Ampliseq, by Life Technologies[63], or Truseq amplicon[64], by Illumina.

**Enrichment by hybridization capture**

Enrichment by hybridization is based on knowledge acquired through microarray research[65] and applied to the enrichment of samples for NGS[66]. Thus, a DNA library is hybridized with probes which represent the target region. Non-specific hybrids are removed by washing and the targeted DNA obtained is eluted.

Two alternative methods exist: array hybridization and in-solution hybridization. The main advantages of both, compared with multiplex PCR are the ability to capture a larger number of targets, and that the coverage obtained is more homogeneous. But the major drawbacks are, firstly, that working with microarrays usually necessitates additional equipment and, secondly, that due to the chemical nature of the process and regardless of the length of the captured region a large amount of starting DNA (10-15μg) is usually required (Figure 04 C)[67, 68].
Solution-based capture overcomes many of these disadvantages: much less DNA is needed, special equipment is not required, and the procedure and is more scalable. The overall performance of solution-based enrichment has also been reported to be better (Figure 04 D) [69].

Roche Nimblegen was the first company to commercialize hybridization enrichment methods, initially array-based, with capture regions of about 4-5Mb. Later, Agilent Technologies also commercialized a similar approach, with similar performance. Currently, the major commercializers are still Agilent (SureSelect[70]) and Roche Nimblegen (SeqCapEZ[71]), but a large number of protocols and vendors are also beginning to offer new capture options, for example Flexgen[72] and MYcroArray[73]. Applied Biosystems and Illumina (TargetSeq[74] and TruSeq[64], respectively) are also offering protocols integrated into the sample preparation platform workflow.

**Enrichment by circularization**

Capture by DNA circularization consists in the use of molecular inversion probes (MIP)[62] or selector probes[75]. It is based on the use of DNA probes, which are single-stranded oligonucleotides consisting of a common linker flanked by target-specific sequences. In MIPs the probes are annealed to the initial DNA sample, undergoing circularization in the process. The hybridized ends of the probe flank the target region (Figure 04 E). The gap between the two ends is then filled and non-circularized DNA is removed by exonucleases.

One specific application of enrichment by circularization, which differs from the previous approaches in terms of hybridising with target regions, lies in Selector Probes,. These probes are designed to accommodate the restriction pattern of the target region (Figure 04 F)[76]. Using these, there is no need to perform downstream steps of library preparation, because the sequencing information for NGS application can be included in the probe[77], which reduces processing time.

Initially, despite being very specific, this approach had low coverage uniformity due to inefficiencies in the capturing reaction itself[47]. Due to implementations mainly by the company Halo genomics[78], acquired by Agilent Technologies, enrichment by circularization is now being commercialized.

**Clonal Amplification**

After library preparation, NGS technologies need to clonally amplify the sample prior to its being sequenced, due to the need for an amplification of the signal generated in order to enable its detection during the sequencing process. This process aims to produce several copies of each DNA library fragment, but must also keep them separated in order to produce an unambiguous, monoclonal signal in the sequencer[39, 42]. Several *in vitro* techniques for clonal amplification have been developed.

**Bridge-PCR**

Bridge-PCR, also called two-dimensional PCR, is a variant of traditional PCR, developed during the 1990s and used first by Solexa[79] in its NGS platforms. Currently this clonal amplification approach is used by Illumina, which acquired Solexa in 2008.

Bridge amplification allows clonal amplification of a large number of DNA fragments simultaneously, using a solid oligonucleotide-coated surface (known as a ‘flow cell’) (Figure 05 A)[80]. The oligonucleotides are complementary to the linkers
added to the DNA sample during library construction. The flow cell is a multi-channel (currently up to 8) sealed glass device. A different library can be added to each channel or the same one can be used in each, depending on the size of the region to be sequenced. DNA fragments from the library are denatured and then attached to the flow cell surface. Oligonucleotides of the flow cell are linked to the surface at its 5’end, leaving the 3’ end free for the polymerase to act on. The resulting double-stranded-DNA is covalently attached to the surface. This double-stranded DNA is then denatured (Figure 05 A1) and the single strand flips over to hybridize to adjacent primers, thus forming a bridge (Figure 05 A2). This newly formed amplicon is extended by polymerases to form a double-stranded bridge (Figure 05 A3). After denaturing, two copies of covalently bound single-stranded templates are obtained (Figure 05 A4)[81]. This cyclical process is repeated several times, producing approximately one million clonal copies of each initial fragment in the form of clonal clusters) (Figure 05 A5). No primers are required in the reaction solution and clusters are spatially separated.

The separation between clusters depends directly in the initial quantity of DNA in the reaction. Higher concentrations allow larger numbers of clusters to be obtained, but also carry the risk of clusters being so close to each other that they interfere with each other. On the other hand, smaller amounts of DNA lead to a lower PCR yield[82].

After bridge amplification, clusters are denatured to give single strands. The sample is then ready for subsequent sequencing steps (Figure 05 C).

**Emulsion PCR**

Emulsion PCR (emPCR) is a variant of conventional PCR that was introduced into commercial use first by 454 Life Sciences, in the 454 sequencer. [83] Subsequent technologies have included it in their protocols, e.g. Applied Biosystems[84] in the SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform[85], Ion Torrent Personal Genome Machine (PGM), and Proton. This technology is also used in the open-source NGS platform Polonator[85].

emPCR is an elaboration of the classical amplification reaction, in which each DNA fragment is isolated in independent aqueous micro-reactors surrounded by an oil phase, and templates are amplified on primer-coated beads (Figure 05 B). An emPCR complex therefore contains millions of compartments separated from each other and each acting as microreactors, in which an independent single-molecule PCR reaction is performed. For its application to NGS, emPCR is usually carried out with beads carrying reaction primers on their surfaces, in order to colocalise the clonally amplified fragments of each reaction on the surface of one bead (Figure 05 B1)[86, 87]. After the reaction, the emulsion is broken down with organic solvents and the beads are isolated by extraction of the aqueous phase.

Each bead will ideally contain multiple copies of a unique monoclonal fragment. However, depending on the amount of DNA added initially to the reaction, there might be more than one DNA fragment in each of the microreactors, leading to generation of polyclonal beads. These are unsuitable for obtaining specific base-interrogation signals, and thus for sequencing. Beads with no DNA fragment amplified on their surface can be eliminated via ‘bead enrichment’. This process consists of the physical separation of amplified and non-amplified beads (Figure 05 B2)[88]. Polyclonal beads, however, can not be eliminated using enrichment methods, but only by limiting the amount of DNA added initially to the reaction.
In subsequent steps, beads are deposited on a solid surface (Figure 05 B3) prior to sequencing (Figure 05 C)[89-91].

**Sequencing Chemistries: Platforms**

There are diverse chemistries commercially available. Despite the differences between them, they all rely in the same principles: base interrogation is performed on all DNA fragments cyclically and in parallel. Signal detection is performed by an optical system composed by a microscope with a Charge-Coupled-Device (CCD) camera plus a computer and storage system, or, in the case of semiconductor sequencing, by a semiconductor chip. Read length of these technologies are usually short, ranging between 35 and 700bp, being the most usual currently 100bp.

**454 Pyrosequencing**

Developed in 1996 by the Stockholm Royal Institute of Technology [92] and commercialized for NGS in 2005 by 454 Life Sciences[89] (currently part of Roche Diagnostics), pyrosequencing is based on detection and quantification of DNA polymerase activity, which is carried out using the enzyme luciferase. Thus, pyrosequencing can be considered a sequencing-by-synthesis method.

The library is first attached to beads via emulsion PCR and then deposited in a PicoTiterPlate which contains millions of wells of 44μm diameter. Each well fits only one bead, of diameter 28μm (Figure 06 A), although not all wells contain a bead[83].

To start the sequencing reaction, a universal primer (complementary to one of the library adaptors), together with DNA polymerase, ATP sulfurlilase, luciferase and apyrase are added to the plate. Then, a modified variant of ATP, which acts as a substrate for luciferase) is added to the reaction. This binds to the complementary 3' positions adjacent to the universal primer previously added to the reaction. The union is catalysed by DNA polymerase and generates one free pyrophosphate per base added. ATP sulfurlilase then uses the released pyrophosphate to generate ATP by combining it with adenosine 5' phosphosulfate. The ATP is then used by the luciferase to generate visible light with an intensity equivalent to the quantity of bases incorporated (and thus available ATP) (Figure 06 B). After each ATP-3' binding and light-generating step, imaging takes place. This allows discrimination of the deposited library fragments that having a base complementary to any of those added. Apyrase then degrades the non-incorporated nucleotides, permitting another reaction cycle to begin, with incorporation of another dideoxynucleotide[93].

The data output of this sequencing platform is not as high as other approaches, having a yield of 700 Mb of sequence per run (23 hours), approximately one million reads and a consensus accuracy at 15x coverage of 99.997%. The relatively low output is compensated for by the high read length obtained (up to 1000 bases) (Table 01)[39, 83]. This read length is useful for performing de novo sequencing on unknown genomes due to its capacity to align complex regions and relatively low computational power that that requires. Moreover, it is possible to sequence up to 132 samples per run with the use of barcodes[94].

**Qiagen-Intelligent Bio-Systems Sequencing-by-Synthesis**

Founded in 2005 at Columbia University (New York), Intelligent Bio-Systems (IBS) has developed its own sequencing-by-synthesis approach. Template preparation is based either on polony amplification or emulsion PCR.

The sequencing chemistry is very similar to Illumina’s approach (see below), except that only a small fraction of the total nucleotides added to the reaction are
labeled (1:100), which lower the costs of reagents while still detecting signal.

The IBS platform, known as Mini-20, claims to be able of sequence 20 samples in 20 individual flowcells, with a maximum output of 20Mb reads per flowcell and reads of 2x100 (80Gb of total output), in approximately 2 days[95]. A larger scale version of Mini-20, MAX-seq, has also been commercialized, although its status has been unclear since IBS was aquired by Qiagen in 2012[96].

While the above IBS platforms have not been widely used, Qiagen’s GeneReader has recently reappeared at the Advances in Genome Biology & Technology Meeting (AGBT, Florida, 2013).

**Illumina Sequencing-by-Synthesis**

Illumina’s sequencing-by-synthesis method is performed after a bridge-PCR amplification process. Sequencing takes place on the solid surface of the flowcell (Figure 07 A). Sequencing-by-synthesis consists of the polymerase-catalyzed addition of reverse-terminator fluorescently labeled bases. Bases are added simultaneously to the reaction and compete to form a union with oligo-primed cluster fragments (Figure 07 B). Once a base is added, it prevents addition of subsequent bases, meaning that only one base will be attached in any one cycle.

An imaging step is performed after base incorporation, to record cluster-specific fluorescence (Figure 07 C). Each flowcell lane is divided in ‘panels’ or ‘tiles’, for a given cluster density (several thousand). Each image represents one tile. After each imaging step, 3’ blocking is chemically removed and the process is restarted (Figure 07 D and E)[17, 39, 97].

This technology was developed in the early 2000s by the US company Solexa, becoming commercially available in 2006. Later, as Solexa became part of Illumina, the chemistry became part of their platforms. Currently, the maximum output available with this NGS sequencing chemistry is approximately 600 Gb per flowcell with 6 billion reads (read length 2x100 bases on a paired-end library) and an accuracy of 99.9%, in approximately 11 days (Table 01); there is also the option of using barcodes (up to 96 samples)[98].

**Polony Sequencing-Danaher Motion Polonator**

Polony (polymerase colony) sequencing is an open-source sequencing chemistry developed at Harvard Medical School[99]. Protocols and software are freely accessible. The open-source implementation of Polony sequencing is the Danaher Motion Polonator G007[90], but it has fallen increasingly into dis-use.

The protocol begins with the preparation of a DNA library (in this case a mate-paired library) and template amplification via emulsion-PCR[100].

Sequencing chemistry[101] begins by attaching a universal ‘anchor’ primer which is complementary to one of the library adaptors and leaving a free 3’OH or 5’P end. A ligase then attaches a base-specific fluorescently labeled nonamer in the position adjacent to that primer. The nonamer bases are degenerate, with the exception of one, whose position will be the one interrogated in the particular ligation cycle. Fluorescence is emitted and recorded by the imaging system after each ligation cycle. The universal anchor primer and nonamer are then separated from the DNA and a new sequencing cycle begins, with addition of an identical universal anchor primer and another nonamer. The position of the non-degenerated base changes each time, so the interrogated position is also different. The result is the generation of millions of 26 bases-reads (2x13 bases from each tag).
Due to its open-source nature, Polony sequencing is a very flexible technique with variable applications, protocols and reagents. However, while a lot of raw data are produced, only a small proportion are useful due to limitations in performance of the technique (Table 01)[85, 102].

It is noteworthy that Polony ligase sequencing has been of significant importance in establishing the basis of other sequencing chemistries, including SOLiD sequencing [103].

**Sequencing-by-Ligation: SOLiD**

Sequencing-by-ligation has its beginnings and basis in older sequencing strategies such as Massively Parallel Signature Sequencing[104] and the aforementioned Polony sequencing[85, 101, 102], and has been commercially available through Applied Biosystems, as the SOLiD system, since 2008.

In this technology, an emulsion-PCR is performed to generate clonal amplicons on the surface of beads, starting from a previously constructed DNA library. Beads are enriched and attached to a glass surface in a random pattern (according to a Poisson distribution) and forming a dense array on which sequencing—by-ligation is then performed (Figure 08 A). First, a universal primer is attached to one of the library adaptors. Fluorescently labeled di-base probes of eight nucleotides in length compete for ligase binding to the DNA fragment adjacent to the universal primer (3' or 5', depending on the tag being sequenced). Di-base probes interrogate the two first bases simultaneously in each ligation cycle. The remaining base probes are degenerate. After ligation, the three final bases of the probe are cleaved so that the length of the probe is reduced to five nucleotides and di-base-specific fluorescence is emitted. After each ligation-cleaving step, an imaging step is performed to detect fluorescence. As in other NGS technologies, the glass array is divided into panels that are each represented as an image. After imaging, a new di-base probe is added and a new ligation cycle begun. By this process, the first two bases of each group of five are interrogated, at three-base interval.

After many cycles of ligation, the extended product and the universal primer are cleaved from the DNA fragment, thus resetting the sequencing reaction. A new universal primer, displaced one base towards the adaptor, is added, such that the sequence interrogation performed by the di-base probes are correspondingly displaced by one base[105]. This process of reset is repeated five times with five different universal primers (Figure 08 B). Finally, every base on each clonally amplified fragment will have been interrogated twice in independent ligations[39, 103]. This two-base encoding system increases sequencing accuracy (Figure 08 C).

Recently, an optional sequencing chemistry add-on has been commercialized, introducing a kind of third-base encoding based on the addition of new probes in which certain positions are already known. This module increases accuracy up to 99.99%. A SOLiD System sequencing run takes 7-14 days (depending on the chemistry used), and currently has a maximum output of 300 Gb per run with approximately 5 billion reads. The read length is 35-75bp with a paired end library (Table 01), and up to 96 barcoded samples can be used[106].

**Ion-Torrent Semiconductor Sequencing**

Semiconductor sequencing can be considered a variant of pyrosequencing in which the detection entity is hydrogen ions generated after dNTP insertion, rather than light. Detection is performed by a semiconductor chip. The concept has been developed by Ion Torrent Systems Inc.[107], which since 2010 has been part of
LifeTechnologies Corporation[91]. The aim of this technology is to enable faster sequencing times by omitting time-consuming imaging steps.

There are currently two Ion Torrent platforms that use this technology: the Personal Genome Machine (PGM) and the Ion Proton. The main differences between these are the size of the sample-deposition surface and the distance between wells, which together determine the sequencing power of the platform. In different platforms there are also different capacities, depending on the size of the chip used. Ion PGM currently has a maximum output of one gigabase of raw data in 2-3 hours, generating up to five and a half million reads with an average read length of 200bp. Ion Proton has announced a maximum raw data output of a hundred gigabases in 24 hours, generating more than three hundred million reads with an average read length of 200bp (Table 01)[107]. Therefore, Ion PGM is focused mainly on the targeted resequencing of gene panels, while Ion Proton is geared towards sequencing exomes and whole genomes.

Clonal amplification of the library is performed through an emulsion-PCR with non-paramagnetic beads. The product of the reaction is enriched (beads that have no amplified product attached are removed) and then deposited on a microwell surface, with each microwell able to accommodate only one bead and positioned above a sensor plate sensitive to changes in pH (Figure 09 A).

To perform the sequencing process, first, an unmodified deoxynucleotide (A, T, G, or C) is flooded into the microwells. Its union with a DNA fragment by a polymerase causes the release of a hydrogen ion, which leads to a pH change inside the microwell. This pH alteration is detected on the sensor plate (ISFET: Ion-Sensitive field-effect transistor[108]). Ion release is proportional to the number of bases added and thus will be the pH change (Figure 09 B).

The main advantages of this sequencing technology lie in the absence of imaging system which is usually expensive and for which data acquisition is time-consuming in the data acquisition process, and in the ability to use unmodified deoxynucleotides. This, allows the attachment of more than one base in the case of repetitive regions. The signal generated does not increase perfectly linear with the number of bases added. As in the case of pyrosequencing technology, that makes it difficult to sequence correctly homopolymers longer than seven bases[109].

**DNA nanoball sequencing: complete genomics**

DNA nanoball sequencing was developed by Complete Genomics (recently acquired by the Beijing Genomics Institute[110]) and released commercially in early 2009. It is specifically designed to sequence whole genomes[111]. This technology consists of the construction of large numbers of compact DNA nanoballs for sequencing. The processing of the DNA sample and library construction differs slightly from the other NGS technologies described above.

Initially, the DNA sample is fragmented and selected to a 400-500bp size. After that, universal adaptors (known as Ad1L and Ad1R) are attached to both ends of the DNA fragments. The sample is then amplified with Ad1-complementary primers and circularized. The product is processed with a restriction enzyme which cuts 13pb internally from Ad1R, making the product linear again. A second set of universal adaptors (Ad2L and Ad2R) is added to the reaction; these also attach to the ends of the DNA fragments, and the sample is again amplified with Ad2-complementary primers and circularized. The product is again processed with a restriction enzyme, this time cutting 13pb internally from Ad2L. After this, a third set of adaptors (Ad3L and Ad3R) is added, and PCR and circularization repeated. This time, the restriction
enzyme is EcoP15, which cuts DNA at 26bp internally to Ad3L, and 26bp from Ad2R. Thus, a fragment of approximately 350bp is eliminated (depending on the size of the initial insert). Finally, the last adaptor set is added (Ad4L and Ad4R), a PCR is performed, and the products circularized, but they are not linearized as before. The sample is then amplified by circular replication using DNA Phi29 polymerase[112]. The DNA strand that emerges from the circular template is a continuous copy (in the same strand) of that template. Due to that the adaptors added to process, the DNA (Ad1, Ad2, Ad3, and Ad4) possesses palindromic sequences and the new strand will collapse onto itself, forming a DNA “ball” of approximately 300 nm (nanoball). Nanoballs are unable to bind to each other by base complementarity. For base-reading DNA nanoballs are deposited in an ordered pattern onto a slide (in contrast to other NGS technologies in which this process is performed in a random pattern) [113].

Sequencing itself is performed using ‘anchor’ oligonucleotides, which are complementary to each end of the Ad1, Ad2, Ad3 or Ad4 adaptors (the technique is therefore also known as ‘combinatorial probe-anchor ligation, or cPAL™ technology [114]). Next, a ligase T4 and a pool of 10mer DNA sequences with degenerate nucleotides in all but the interrogative position are added to the reaction. The base interrogated with the 10mer probe is correlated with the fluorophore attached to that probe. A probe will only bind to the DNA if it is complementary in the interrogated position. Non-binding probes are washed away and, fluorescence is detected in an imaging step. After each probe reading and base interrogation, the used anchor and 10mer are removed, and new ones added. The process is repeated until all ten interrogative positions next to the anchor sequence have been recorded. Once that has been achieved, a new anchor that binds to a different adaptor end is added and the process of interrogation repeated. A total of 70 bases can be sequenced using this technology. Due to template processing with Ad1, Ad2, Ad3 and Ad4 adaptors, the bases sequenced are the first and last 35 with respect to the initial DNA size-selected fragment of 400-500bp.

The main advantages of this technology with respect to other sequencing platforms are: 1) The use of very high-density arrays by nanoball dispersion in an ordered gap-free pattern; and 2) the non-progressive nature of the sequencing. Having been designed specifically to sequence whole genomes, nanoball sequencing also claims to be cheaper than other NGS platforms for that specific application. Disadvantages to consider are the short read length of sequences obtained, which complicate the bioinformatic process of aligning the sequence with the reference genome, and the multiple rounds of PCR necessary[115, 116].

The Complete Genomics sequencing platform produces 60 gigabases of usable data from each slide in a 12-day run - therefore to sequence a complete genome with 40x coverage, 120 usable gigabases are needed. 18 slides can be processed at a time, and accuracy is said to be up to 99.999% for 90x coverage (Table 01).

**NGS Bioinformatics**

The development of NGS platforms makes huge demands on statistical methods and bioinformatics tools for the analysis and management of the data generated. Indeed, bioinformatics analysis has become one of the main bottlenecks for NGS. As described in previous sections, the data throughput of these platforms has increased by two orders of magnitude compared with older technologies. It is therefore necessary to continue to develop informatics algorithms to translate the results of DNA sequencing into bites of information that can be managed by researchers, physicians, patients, and others[117]. Such algorithms can be organized
into several categories: 1) alignment of sequence reads to a reference; 2) base-calling and/or polymorphism detection\cite{118, 119}; 3) de novo assembly; and 4) genomic viewers\cite{120-125}.

Alignment algorithms have to allow for the fact that the reads obtained with these technologies are short; widely used solutions such as BLAST, BLAT, or CLUSTALW do not therefore constitute the best options. A large number of new algorithms specifically designed to manage short reads have now been developed or adapted from older algorithms\cite{126, 127}. These new tools outperform the previously mentioned traditional aligners. In recent years, several dozen short-read alignments have been published, example Maq\cite{119}, Bowtie\cite{128}, SSAHA\cite{127}, BWA\cite{129}, SOAP2\cite{130}. Short unpaired reads represent a particular challenge for de novo assembly\cite{131, 132}, with mate-paired reads being the most appropriate for that purpose\cite{133}. Alignment software has to take into account the estimated quality of the underlying data in order to make alignments correctly\cite{119}. Many new algorithms have also been developed for sequencing de novo: AbySS\cite{134}, ALLPATHS\cite{131}, Edena\cite{133, 135}, SOAPdenovo\cite{126}, Velvet\cite{134}. Challenges still remain in the reconstruction of repetitive regions.

A handful of tools have been developed for base-calling and polymorphism detection, which are important applications of NGS technologies. Some tools filter reads based on the initial alignment quality, such that the conditional likelihood of nucleotides at each position can be computed, following a bayesian approach and giving the probability of a particular variation to be real. Correct annotation of the variant found is essential to correctly understand its function. Many tools\cite{136-140} and databases\cite{4} are also being used and developed to carry out fast filtering of functionally important variants, with a specific language and ontology\cite{141}. A complete list of algorithms and software is available online\cite{142}.

For RNA sequencing, some additional aspects have to be considered, to meet challenges such as transcriptome sequencing, RNA splicing, and the extremely short lengths of reads in small RNAs. In contrast to other approaches, therefore, alignment algorithms have, not only to map reads to the genome, but also to take into account the possible presence of splice junctions and apply algorithms to determine and quantify the most likely splicing isoform\cite{143, 144}.

One major type of algorithm for RNA sequence alignment comprises unspliced read aligners. These do not allow gaps between exons and can be divided into two sub-categories: Seed methods (MAQ\cite{119} and Stampy\cite{145}) and Burrows-Wheeler transformation methods (BWA\cite{129} and Bowtie\cite{128}). The other major type of algorithms consists of spliced aligners, which do allow gaps between exons and can be further divided into: exon-first (MapSplice\cite{146}, SpliceMap\cite{147} and TopHat\cite{148}) and seed-extend (GSNAP\cite{149} and QPALMA\cite{150}). For transcriptome reconstruction against the reference genome, the process can be either genome-guided or be genome-independent.

For Chip-Seq, the analytical requirements are distinct from the previously described approaches, because it is important to correctly calculate the confidence of the 'peak' of the binding domain. Common tools for this analysis are the PeakSeq tool\cite{151} and the SPP tool\cite{152}.

For data visualization, a genome browser is required; this is a graphical interface designed to show NGS genomic data (reads) aligned to a reference genome. Such browsers usually allow the acquisition of annotated data from databases (dbSNP, 1000Genomes, etc), and thus comparative analysis, prediction of
pathogenicity of variants, expression changes, or protein folding. Such tools include Eagleview, Mapview, Maqview, Tablet, and IGV.

It is also important to consider the huge computational power needed to align the millions of RNA reads generated by a standard platform, which necessitate large dedicated centers or cloud computing. Data storage is also an important consideration, with the sequencing cost per base dropping faster than the cost of storing a byte on a hard disc. NGS throughput, in fact, has overcome Moore’s Law of Informatics[153].

Applications of Next-Generation Sequencing

In recent years, the number of publications on NGS have increased substantially, demonstrating its capabilities in many areas and to many applications. Important applications include: whole genome resequencing[154], targeted resequencing[155], de novo sequencing[156], gene expression analysis with whole transcriptome analysis[157], small RNA sequencing[158], methylation analysis[159], the sequencing of DNA extracted by antibodies targeting DNA-binding proteins in proteome research (ChIP-Seq)[160], and nuclease fragmentation and sequencing. These applications are currently used to conduct studies in a large number of areas: 1) genetic diagnosis of common and rare diseases and the discovery of causal genes by targeted resequencing[20, 53, 161-170]; 2) cancer research[22, 35, 171-175]; 3) microbiological studies[7, 156, 158]; 4) evolutionary and population studies, including the use of ancient DNA[21, 33, 36, 176-178]; 5) prenatal diagnosis from maternal blood samples[179]; 6) transcriptome studies[157]; 7) personal identification/forensics[8]; 8) metagenomic studies[159, 180]; 9) mitochondrial genome studies[177, 181]; 10) immunology studies[170]; and 11) plant biology[10].

The ultimate aim of all these applications, as for any gene-mapping technology, is to understand the functional consequences of DNA sequence variation and how that leads to disease. There is the additional difficulty of our limited understanding of non-coding regions of the genome, which previous methods could not address. Now, due to worldwide projects such as ENCODE, this information about the functional organization of the genome is beginning to be uncovered[182, 183].

Due to the affordable cost and high sensibility of NGS, there is currently an increase in the use of exome-sequencing. This is especially useful in family inheritance analysis, to find a causative gene in rare Mendelian disorders. NGS allows the laborious process of searching for a causal mutation via linkage analysis and fine-mapping by Sanger sequencing to be bypassed. This has led to the discovery of causative genes in many diseases in which causal alleles disrupt protein-coding (exonic) sequences[12, 184-186].

There is also the need to analyze multiple genes that show overlap in structure and clinical phenotype, without having to sequence the entire exome. Customizable genomic enrichment techniques mentioned in previous sections offer such an approach. There are in fact many commercially available options, such as predefined gene panels, to study diverse types of related diseases or groups of genes (cancer, inherited diseases, the kinome)[20]. It would, however, take several decades to identify the underlying factors of many complex diseases, since they are influenced by many subtle environmental and genetic factors[187].

Such problems are often viewed as an impasse to the ability to perform whole-genome sequencing at an affordable cost. As complex diseases are driven by both
coding and non-coding variation, the dataset needed for its study comprises the complete sequence of the genome of an individual and its comparison with databases containing thousands of genomes. In principle, this would make it possible to compare all types of DNA variation (SNP, indels), frequencies, and functions (coding, non-coding). While being very powerful, such a project would also be extremely complex due to the added difficulties inherent in sequencing repetitive or degenerated regions of the genome. Due to the relatively short read-lengths of NGS technologies, such regions are still excluded. Using the complete sequence of the human genome, NGS chemistries can currently make alignments of approximately 85% against the reference sequence (which is increased to 93% if only coding sequences are considered). Longer reads will improve these values[4]. It is also possible to perform whole-genome haplotyping of family genome sequences, using appropriate analysis algorithms[188]. The extra information can be used to increase the power of analysis in disease-association studies, recombination studies, allelic exclusion, and in population genetics (migration studies and evolutionary selection). Finally, such information can facilitate the discovery of modifying genes in Mendelian disease, together with their assignment to a haplotype. The only weakness of this approach is that it is limited to individuals who have access to the genomic sequences of their parents and siblings. There is also the possibility, described by Peters et al [189], to perform this kind of haplotyping using a very small number of cells (1-20) as the starting material.

The high sensitivity of these techniques is giving new research insights that, with time, will not only increase the resolution of routine diagnostic procedures but reveal new dimensions in diseases previously classified exclusively by traditional parameters (morphology, pathophysiology, etc)[169] and lead to the development of new diagnostic methodologies. It is, for example, possible for the first time to track the evolution of a cancer by sequencing it at many developmental stages, by performing micro-biopsies and whole genome sequencing and haplotyping [189], or exome sequencing [190, 191]. It is even possible to detect mutations in tumor cells directly from the circulation, or to sequence cell-free fetal nucleic acids (DNAs and RNAs) from maternal blood during pregnancy[13, 192]. Several approaches using this latter technique have been developed to determine fetal aneuploidy, by counting the number of reads that map to each chromosome and to calculate the relative dosage of each.

Transcriptome sequencing, on the other hand, is the complete sequencing of all the transcripts in a cell, and their quantification at any specific developmental stage; it is essential for interpreting the functional elements of the genome and for tracking both development and disease. Transcriptome sequencing studies have been used for applications ranging from gene-expression profiling across many cell types and/or conditions, genome annotation, and non-coding RNA discovery and quantification. The most usual application is the sequencing of poly(A)-containing RNA molecules, principally the messenger RNA fraction of the transcriptome[157]. Alternative protocols include strand-specific RNA-seq, and sequencing of poly(A)-negative RNA. The analysis of small RNA (transfer RNA, ribosomal RNA, small nuclear, small nucleolar, microRNA and small interfering RNA) is also becoming an important application of these technologies, as they produce appropriate read lengths for it. High-throughput sequencing of small RNAs provides great potential for the identification of novel small RNAs and for profiling those that are already known.
NGS can also be applied to pathogen detection via small RNA sequencing[158], or to metagenomic-based strategies used for the discovery of novel microorganisms and characterization of microbial communities. Because it is possible to study a complete microbial population, there is no longer a need to isolate and culture individual microbial species, and prior knowledge of their sequences is not necessary[193]. It is interesting also to note the versatility of the data, which can be analyzed to provide insight into the level of gene expression, structure of genomic loci and sequence variation.

In mitochondrial disorders, NGS can be used to sequence an entire mitochondrial genome (16.5kb), determine percentage mutation heteroplasmy, and analyze mitochondrial-related nuclear genes, all in a single run[194].

The association between DNA and proteins is also a fundamental biological phenomenon which plays a key part in the regulation of gene expression and replication, as well as many other processes. In this context, hromatin immunoprecipitation can also be adapted for investigation using NGS. This has been demonstrated by Johnson et al [195], who have generated a genome-wide, highly precise positional profile of chromatin following degradation of the genomic DNA by nucleases. Two of the main areas of interest for Chip-sequencing is the study of transcription factor binding sites[196, 197] and in profiling histone modification [198].

However, in spite of the very wide range of possible applications and uses for NGS technologies, together with the advances being made in this area, their transition into routine clinical diagnosis is still in the early stages of development. Undoubtedly, this is partly due to the enormous computational requirements and the vast amount of data needing analysis.

**Third Generation Sequencing: an insight into the near future**

While in many respects NGS is now in full swing, other technologies, known as ‘Third Generation Sequencing technologies’ (TGS) are starting to appear. The main advantage of TGS is their ability to sequence single molecules of DNA[199] with no need for clonal amplification prior to sequencing. This avoids the introduction of artefacts from PCR and requires less manipulation of the sample in comparison with NGS. TGS also usually involve sequencing-by-synthesis chemistries, but detection techniques are based mainly on the physical recognition of DNA bases in an unmodified DNA strand, rather than on detection of chemical incorporation[17]. In these newer technologies, moreover, the sequencing reaction is not paused for ‘wash and scan’ steps after the incorporation of each base, so time and reagent-consumption steps are avoided[200]. TGS also represents a new challenge for bioinformatics; the data obtained with these technologies is of a different kind compared with SGS or Sanger sequencing, due to their ability to detect kinetic activity in the polymerase and/or the influence of secondary structures on sequencing speed.

This type of sequencing is currently being developed by a large number of companies, each using different chemistries, and can be categorized as follows.

**Single-Molecule Real-Time Sequencing**

Developed by Pacific Biosciences[201] (previously known as Nanofluidics Inc), this technology is based on the detection of natural DNA synthesis by a single
DNA polymerase. Specific phosphate-labeled nucleotides (known as phospholinked nucleotides) are added simultaneously, so that the polymerization occurs in a continuous and processive manner[202]. To understand this approach, two main innovations should be considered: the aforementioned phospholinked nucleotides and the Single-Molecule Real-Time Sequencing chip (SMRT), a nanophotonic visualization chamber that contains thousands of zero-mode waveguides (ZMWs)[199, 203]. These latter are cylindrical holes of approximately 45nm diameter in a 100nm thick aluminum film, which together provide a fluorescence detection volume of about 20 zeptoliters. This, confines the enzyme to a tiny volume which increases the signal-to-noise ratio[204]. When phosphate is cleaved as the nucleotide is incorporated into the DNA strand, base-specific fluorescence is released. After incorporation, the phosphate and dye diffuse quickly out of the tiny detection volume, maintaining the background signal at low levels. As the dye is eliminated when phosphate is cleaved, the result of polymerization is a single long, natural DNA strand (Figure 10). The sequencing instrument detects base incorporation as a specific-color flash at a rate of multiple bases per second.

Each SMRT chip has approximately 75000 ZMWs, each with a DNA polymerase loaded with a different strand of DNA sample. Thus, 75000 single-molecule-sequencing reactions occur in parallel, with no time-consuming scanning or manual detection steps (signal detection is performed by video recording). Sequencing runs therefore last minutes or hours instead of days[205].

Average read length is above 1000bp and up to 10kb, enabling easy de novo assembly. Sample preparation is also easy, consisting of fragmenting DNA, blunting the ends and ligating hairpin adaptors. It is also possible to make independent reads of the same molecule via circular consensus sequencing, which increases the read quality of linearly with the number of reads. However, sequence length decreases to 250bp in this case[206].

SMRT sequencing instruments allow applications that are not possible with other technologies, such as real-time enzyme activity observation (e.g. of the ribosome as it translates mRNA, RNA-dependent polymerases, etc)[207]. Also, by observing polymerase activity, kinetic information on nucleotide incorporation can be obtained, which in turn allows data on DNA modification and the structural nature of the strands to be gathered. Thus, the data obtained using TGS are different from those using SGS.

While read accuracy has been reported to be about 83%, this can be improved to more than 99% with 15x coverage. Throughput is also not as high as in SGS[25].

Real-Time DNA Sequencing using Fluorescence Resonance Energy Transfer

This chemistry is still in its early developmental stages and few data are available. In this approach, DNA polymerase with an attached fluorophore is brought into close proximity with a phospholinked nucleotide, which is itself tagged with an ‘acceptor fluorophore’[208]. The interaction causes emission of a fluorescence resonance energy transfer (FRET) signal, and the fluorophore label on the nucleotide is released.

Due to the high speed of the process, sequencing can be performed at millions of bases per second using this technique[200].

Direct Imaging of DNA using Electron Microscopy.

Halcyon Molecular [209], and ZS Genetics[210] are developing this approach. It allows direct imaging and chemical detection of atoms to identify nucleotides[211].
Its ability to detect atoms in a planar surface by annular dark-field imaging in an aberration-corrected scanning (TEM) has recently been demonstrated as proof-of-concept[212]. Halcyon is developing supporting technologies in parallel, to allow TEM DNA sequencing work, as for example the use of functionalized needles to attach stretched molecules of DNA to a substrate for the imaging.

The main drawbacks of this technology is that a one-cycle PCR step is needed, specifically to label each of the four nucleotides with heavy elements to allow their detection if on the planar surface. Even with one cycle, PCR-based modifications can be introduced. Good sample preparation is critical, and there is a limit on speed due to imaging steps[25].

**Direct Imaging of DNA Sequences using Scanning Tunneling Microscope Tips**

There is no proof of-concept yet published by the main companies involved in the development of this approach (Reveo Inc[213]). In this technique, DNA is placed on a conductive surface from which bases are detected by Scanning Tunnelling Microscope Tips (STM Tips). Individual bases will probably be detected by tunnelling current measurements. This is expected to become a cheap and rapid technology due to a lack of labeling steps[200].

**DNA Sequencing with Nanopores**

Nanopore sequencing is one of the most advanced third generation sequencing techniques. It consists of the transit of a DNA molecule or its component bases through a pore[214] and its sequence interrogation due to their effect on an electric current or optical signal[215, 216]. The system is contained in a synthetic lipid bilayer such that when DNA is loaded onto one face and a voltage applied across the bilayer (by changing the concentration of salt), DNA strands can be mobilized through the pore (Figure 11) [200]. One of the most important advantages of nanopore sequencing is its inexpensive sample preparation due to the use of unmodified DNA and the employment of a nanopore sensor which eliminates the need for nucleotides and polymerases or ligases. Very long read lengths are also expected.

Many approaches to nanopore sequencing exist;
1) Strand-sequencing using ionic current blockage through an alpha hemolysin pore.
2) Exonuclease sequencing by modulation of the ionic current. The exonuclease attached to the pore cleaves dNMPs from the end of a strand. Base recognition is determined by the level of current blockade of each nucleotide.
3) Nanopore sequencing using synthetic DNA and optical readout[217].
4) Strand-sequencing using transverse electron currents[218].
5) Sequencing using solid-state graphene nanopores. Research is currently being carried by the Cees Dekker Laboratory (Delft, Holland)[219]. This technique has the advantage of the nanopores being narrower than the distance between two bases of DNA (0.3nm and 0.5nm respectively), and graphene being mechanically robust and a very good electrical conductor. It has already been shown that single molecules of DNA can be pulled though graphene nanopores and simultaneously detected[220, 221].
6) Nanopore sequencing-by-synthesis (Nano-SBS). In this approach, a polymerase is attached to the pore, and nucleotides added to the reaction are specifically tagged. As each nucleotide is incorporated into the template DNA, the tag is released, producing a unique ionic blockade (as tags are of different sizes), thus determining the DNA sequence[222].
Oxford Nanopore Technologies has recently introduced its nanopore sequencing platform, GridION, together with a disposable nanopore sequencing device the size of a USB memory stick, called MinION\cite{223}.

**Transistor-Mediated DNA Sequencing**

IBM is currently developing a nanostructured sequencing device to electronically detect individual bases in a single molecule of DNA. The structure consists of metallic pores with alternating layers of metal and dielectric materials. As DNA passes through the pores, each base is specifically recognized by its specific interaction with the device\cite{224}. Due to the nature of the technique, it is expected to achieve extremely high throughput (millions of bases per second). As other TGS approaches, this one is claimed to be inexpensive due to its label-free optics-free characteristics\cite{225}.

**Conclusions and Future Perspectives**

Revolutionary changes that have occurred in the life sciences coincide temporally with technical advances and major breakthroughs in sequencing technologies, thus demonstrating their utility and value in addressing complex biological questions. The spread and popularization of NGS due to its reduction in cost-per-base is democratizing access to these technologies, with investigators able to carry out projects at scales previously only accessible in large genome centers. This has been accompanied by an increased interest in sequencing, reflected in the growing number of research groups using the technologies in their research and companies working on future technologies. In this respect, TGS still has much to demonstrate in terms of translating technology into useful data. NGS, therefore, has not only been crucial on the beginning to elucidate genome variability, but continues to be critical in advance the understanding of living systems and complex phenotypes.

None of these advances would be possible without also developing appropriate mathematical algorithms to transform the sequence reads and associated data into meaningful information. In many cases algorithms are still being developed. Despite this, due to current NGS technologies major advances are being possible in many areas, especially, as mentioned before, in the fields related to the understanding and diagnosis of rare and complex diseases\cite{226}. NGS technologies therefore constitute the first opportunity for researchers and clinicians to understand the genetic variability underlying such problems, and provide an unprecedented tool in their investigation, as reflected in the year-on-year increase in the number of relevant publications.

In conclusion, in this review we have described the current state of sequencing technologies. While it is difficult to predict even the near future of sequencing, it is clear that NGS will become an increasingly routine tool. Moreover, as our knowledge of variability in the human genome increases so too will our functional annotation of the genome. In turn, this will enable sequencing data to be much more readily interpreted by professionals, in terms of potential impact. Currently one of the major impediments of these technologies in order to be of routine use.
References


[41] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program Available at: www.genome.gov/sequencingcosts. Accessed [2012/08/13].


[56] www.products.invitrogen.com/ivgn/product/A10498


[58] wwwfluidigm.com

[59] www.raindancetech.com


[64] www.illumina.com/truseq-ilmm


[72] www.flexgen.nl

[73] www.myroarray.com/myreads/myreads.html#Capture


[78] www.halogenomics.com


[81] www.illumina.com


[84] www.appliedbiosystems.com


[89] www.454.com
[90] www.polonator.org
[91] www.lifetechnologies.com
[92] www.kth.se/en
[96] www.intelligentbiosystems.com/index-1%20mod%201.html#
[98] www.illumina.com/documents%5Cproducts%5Ctechnotes%5Ctechnote_Q-Scores.pdf
[99] www.hms.harvard.edu
[105] www.solid.appliedbiosystems.com
[107] www.iontorrent.com


[110] www.genomics.cn

[111] www.completegenomics.com


[114] www.completegenomics.com/services/technology/


[120] www.broadinstitute.org/igv

[121] www.bamview.sourceforge.net

[122] www.evolution.sysu.edu.cn/mapview


[137] www.sanger.ac.uk/resources/software/vagrent/

[138] www.pupasuite.bioinfo.cipf.es/

[139] www.genetics.bwh.harvard.edu/pph2/

[140] www.mutationassessor.org
[141] www.sequenceontology.org


[201] www.pacificbiosciences.com


[209] www.halcyonmolecular.com

[210] www.zsgenetics.com


[213] www.reveo.com


[219] www.ceesdekkerlab.tudelft.nl/research/projects/nanopores/


[223] www.nanoporetech.com


Table 01: Compared characteristics of the main NGS platforms.

Figure 01. Schematic representation of Sanger sequencing workflow. A1) The sample is first enzymatically digested and cloned into bacteria. Many copies of a given initial DNA fragment are collected from each bacterial clone. A2) With the appearance of the PCR in mid 80’s the process of DNA amplification were significantly shortened. Millions of copies of the initial target are obtained in a few hours. B) Chain termination method relies on the ability of ddNTP’s to stop chain elongation. The differently-sized products are separated in an electrophoresis. In this schema each ddNTP is fluorescently-labeled. Signal detection can be automated by the use of a laser/computer.

Figure 02. Representation of the drop-down cost per Mb sequenced. Notice that in the last ten years, costs have move down five orders of magnitude, being the cost per Mb since Mid-08 less than 10$.

Figure 03: Representation of the most used DNA libraries configuration for Next Generation Sequencing. A) A fragment library consist in the representation of the initial sample in DNA fragments of variable size (starting in 150pb up to 1000) flanked by universal adaptors. One or two tags of sequence can be obtained per fragment (A1 and A2 respectively). B) A barcoded library includes a indexing, specific sequence in order to multiplex sample preparation and sequencing. Sequencing tags can be single or paired end (B1 and B2). C) A mate paired library consist in the representation of the initial sample in pairs of DNA fragments which separation in the reference sequence is previously known. Sequencing approach depends on the platform used (C1, C2, and C3).

Figure 04: DNA library preparation scheme in current Next Generation Sequencing. * Indicates non mandatory steps depending on the application.

Figure 05: Schema of the current main approaches for clonal array generation.

Figure 06: Pyrosequencing steps. A) The DNA library is attached to beads and deposited in wells on a PicoTiterPlate. B) The objective of pyrosequencing is to measure the light emitted by the luciferase. To link it to the base addition to the growing strand of DNA a number of chemical reactions are necessary. First a known dNTP is added to the reaction and binded by a polymerase to the complementary 3’ positions adjacent to a universal sequencing primer. This, generates one free pyrophosphate per base added. ATP sulfurilase will generate ATP by combining this pyrophosphate with adenosine 5’ phosphosulfate. This ATP will be the luciferase substrate to generate visible light. After light measurement over each bead, the enzyme apyrase will degrade the non-incorporated nucleotides and a new sequencing cycle will begin.
Figure 07: Sequencing by Synthesis steps. A) Sequencing reactions take place in the clonal clusters on the flowcell surface. B) First, bases are added to the reaction, competing for the union to a primer attached to a universal adaptor on each library fragment. Bases are blocked at its 3’end, so only one is added. C) After incorporation, fluorescence is emitted and registered by the imaging system. D) 3’ blocking is removed. E) The previous steps are cyclically repeated in order to sequence the desired number of bases.

Figure 08: SOLiD sequencing by Ligation steps. A) Sequencing is performed on the surface of DNA-coated beads attached to a glass slide (also known as Flowchip). B) First, a universal primer is attached to one of the library adaptors. Di-base probes compete for the union and subsequent ligation to the primer-adjacent position. After ligation, the three final bases of the octamer is cleaved and fluorescence is emitted. After N ligation cycles (usually 10 but may vary) the reaction is resetted and a new displaced primer is added to the reaction, so different bases are interrogated. This process of resetting and primer addition is performed five times per sequencing tag. By this, each base of the sequenced DNA fragments will be interrogated two times in independent reactions. C) This two base encoding system entails the need of decoding the information obtained by the double interrogation. D) Color decoding matrix.

Figure 09: Semiconductor sequencing. A) Sequencing reactions are performed inside a microwell above a sensor plate. B) First, an unmodified, known dNTP is added to the reaction. It is attached to the 3’ end of a sequencing primer by a polymerase. By this, a hydrogen ion is released with each base added. These ions -pH alteration- will be detected by the sensor plate, so no imaging steps are needed. After detection and washing a new deoxynucleotide can be added to the reaction, repeating the process.
Table 1

<table>
<thead>
<tr>
<th>Platform</th>
<th>Clonal amplification</th>
<th>Sequencing chemistry</th>
<th>Read length (bp)</th>
<th>Throughput (Mbp/run)</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>emPCR</td>
<td>Paired end</td>
<td>300-600</td>
<td>600,000</td>
<td>11 days</td>
</tr>
<tr>
<td>Ion Torrent-PGM</td>
<td>emPCR</td>
<td>Synthesis (semiconductor sequencing)</td>
<td>300-600</td>
<td>45,000</td>
<td>2.5 hours</td>
</tr>
<tr>
<td>Illumina</td>
<td>Bridge PCR</td>
<td>Synthesis</td>
<td>150-250 Paired end</td>
<td>150,000</td>
<td>14 days</td>
</tr>
<tr>
<td>Applied Biosystem’s SOLiD</td>
<td>emPCR</td>
<td>Ligation</td>
<td>150-250 Paired end</td>
<td>150,000</td>
<td>14 days</td>
</tr>
<tr>
<td>Nanoball sequencing</td>
<td>Template circular replication</td>
<td>Ligation</td>
<td>150-250 Paired end</td>
<td>150,000</td>
<td>14 days</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>emPCR</td>
<td>Synthesis (semiconductor sequencing)</td>
<td>150-250</td>
<td>150,000</td>
<td>2.5 hours</td>
</tr>
<tr>
<td>Helicos (Helioscope)</td>
<td>None (SMI)</td>
<td>Synthesis</td>
<td>150-250</td>
<td>150,000</td>
<td>2.5 hours</td>
</tr>
</tbody>
</table>
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5

Clonal array generation

A
Bridge PCR

A1
Primed-coated surface
DNA Polymerase

A2

A3

A4

A5

B
Emulsion PCR

B1
Primed-coated bead
DNA Polymerase
Input DNA library
Primers

B2
Bead enrichment

B3
Bead deposition

C
Cyclical parallel base interrogation

Cycle 01
Cycle 2
Cycle n
Figure 6
Figure 7
Figure 8
Figure 10
Nanopore Sequencing

Figure 11
Highlights

Next-generation sequencing is becoming a widespread technique used in many areas. Some initial guidelines are usually needed to take advantage of this approaches. This review aims to provide general knowledge on next-generation sequencing basics.