

# “Next-generation” Sequencing (NGS): The new genomic revolution

## **Berta Fusté**

Unitat Genòmica, CCiTUB, Universitat de Barcelona. Parc Científic de Barcelona. Baldri Reixac, 10. 08028 Barcelona. Spain.

email: [bfuste@ccit.ub.edu](mailto:bfuste@ccit.ub.edu)

**Abstract.** In the past 5 years “Next-generation” Sequencing (NGS) technologies have transformed genomics by delivering fast, inexpensive and accurate genome information changing the way we think about scientific approaches in basic, applied and clinical research. The inexpensive production of large volumes of sequence data is the main advantage over the automated Sanger method, making this new technology useful for many applications. In this chapter, a brief technical review of NGS technologies is given, along with the keys to NGS success and a broad range of applications for NGS technologies.

## 1. Introduction

Nearly three decades ago Fred Sanger [1] and Wally Gilbert [2] received the Nobel prize for chemistry sequencing technology (see Automated electrophoresis Technique in “Genomics: More than genes” in this issue). This technology known as “Sanger Sequencing” or “First-generation sequencing” has dominated the industry for almost two decades and led to a number of accomplishments, including the completion of the human genome sequence. In 2004, the International Human Genome Sequencing Consortium reported a genome sequence with high accuracy and nearly completed coverage of the human genome [3]. The approach consisted in “Sanger sequencing” more than 20,000 large bacterial artificial chromosome clones in which each one contained 100 kb fragments of the human genome. This project was performed during 10 years with multiple groups and companies participating. Although many technical improvements were achieved in this era, like automated capillary electrophoresis among others, new inventions were needed for sequencing large numbers of genomes. Recent efforts have been directed towards new methods known as “Next-generation” sequencing (NGS) technologies, which are capable to sequence a complete human genome in one week. In 2004, the 454 FLX Pyrosequencer from Roche was the first NGS sequencer to become commercially available [4]. The Solexa 1G Genetic Analyzer from Illumina was commercialized in 2006 [5]. The SOLiD System from Applied Biosystems was launched in 2007 [6]. Most recently, a “Third-generation” of sequencers has emerged. The HeliScope from Helicos BioSciences started shipping in early 2008 [7] and PacBio RS from Pacific Biosciences in 2010 [8]. Although the first new technology appeared in 2004 more than 1000 next-generation sequencing-related manuscripts have appeared in the literature. NGS technology is not only changing our sequencing approaches and lowering the associated time and costs, it is also opening a wide variety of applications, types of biological samples, and new areas of biological inquiry.

BT.7

## 2. Methodology

### 2.1. Principles of NGS technology

The principle of the NGS technologies can be summarized as the sequencing of a multiple-parallel array of DNA features using iterative cycles of enzymatic manipulation and imaging-based data collection [9]. Although the different platforms are quite diverse in sequencing biochemistry and generating the array, their work flow is conceptually similar. The first step consists in Library construction which is achieved by random fragmentation of DNA, followed by an *in vitro* ligation of common adaptor sequences. The second step implies the generation of clonally clustered amplicons to serve as sequencing templates. Polymerase Chain Reaction (PCR) amplicons derived from a single library molecule end up spatially clustered, either to a single location on a planar substrate (bridge PCR [10, 11], or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR [12]). Subsequently, the sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition (Figure 1). Most of the platforms use sequencing by synthesis, that is, serial extension of primed templates, but the enzyme driving the synthesis can be either a polymerase [13] or a ligase [14]. Data are acquired by imaging of the full array at each cycle.

Besides their technical differences, other main concepts differ between platforms are: length of the generated sequences, production capacity that each platform can achieve per run, real-time DNA sequencing, and no need to amplify the sample before sequencing. These important points can lead to choosing one platform or other depending on the organism, application or expected result. Among the NGS platforms, the FLX-GS from Roche is the only one capable to generate long reads (average fragment length is around 400bp) but its sequencing capacity is low (350-400Mb). On the contrary, other platforms get their strength in their sequencing capacity. HighSeq 2000 from Illumina generates up to 200Gb per run and 5050XL SOLiD system produces up to 20-30 Gb per day with the inconvenience that their fragment length is among 50-100bp. This is an important issue in order to determine which platform is optimal for your experiment. For example,

if a biological sample is an unknown genome, longer reads are needed to assemble it, but if your sample has a reference genome available and the study consists in variation discovery in individuals, high sequencing capacity plays a key role. On the other hand, creating great expectations, and already in the final experimental stage, are the “Third-generation” sequencers. This is the case of the new Real-time sequencing machine from Pacific BioSciences, which harnesses the natural activity of key enzymes involved in the synthesis and regulation of DNA, RNA or proteins. Also, there is the HelicoScope, which ADN amplification is not needed before sequencing, this being really interesting to avoid amplification bias, although in the present its fragment average-length is too short.

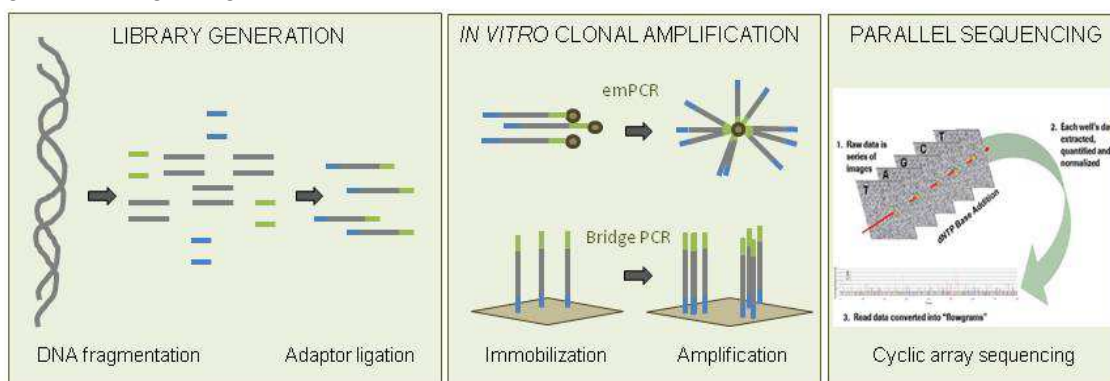


Figure 1: Work flow of Next Generation of DNA Sequencing.

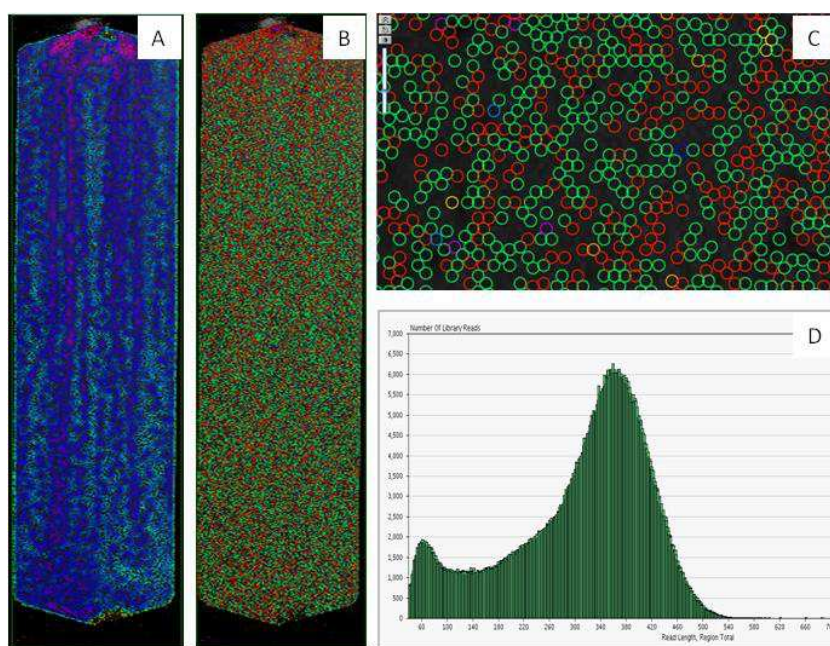
BT.7

## 2.2 Keys to NGS success: NGS vs. Sanger

Under standard conditions, Sanger sequencing results in a read length between 700-1000 bases. These are relatively long read lengths compared with NGS methods (400bp-Roche, 75bp-Solexa and 50bp-SOLiD). However, first generation sequencing is limited by the small amounts of data that can be processed per unit of time. The major advance offered by NGS is the ability to produce an enormous volume of data inexpensively [15]. In fact, the first version of the 454's instrument could easily generate a throughput equivalent to that, or more than 50 Applied Biosystem's 3730XL at about one-sixth of the cost [16]. The array based sequencing approach, where hundreds of millions of reads are obtained in parallel on a small area, lowers the cost of DNA sequence production since the same reagent volume is used for all the sequences at once (Figure 2).

Another important point is the *in vitro* construction of a sequencing library, followed by *in vitro* clonal amplification to generate sequencing features, avoiding “*E. coli*” transformation and colony picking. Both steps are possible thanks to the common adaptors ligated in the library construction. These adaptors are also the responsible to extremely reduce the time, and consequently the cost of sequencing.

The accuracy, the degree to which a reconstructed sequence is free from errors, is one of the topics that can be considered as a disadvantage of NGS vs. Sanger sequencing although the greater output overcomes this challenge. NGS platforms can produce high number of reads providing greater depth and sequence confidence. Controversially, this great output from NGS technology comes with data management and analysis challenges. Significant computational infrastructure is required to process raw data generated by the sequencers and assemble or align them into contigs or scaffolds. But not only the computational infrastructure is a problem, a limited set of applications are covered with the instrument manufacturers software, and the need of specialized bioinformaticians is crucial for the success of this technology (see “Bioinformatics: making large-scale science happen” in this issue).



**Figure 2.** Results obtained from a run performed in a GS-FLX instrument from Roche at the Genomic Unit, CCiTUB. The fluidics subsystem ensures accurate reagent dispensing and flows the sequencing reagents across the wells of the PicoTiterPlate device. Panel A shows raw density of the beads load in a PicoTiterPlate. Panel B represents the quality control of all the reads sequenced in the same run. Panel C is an augmented area of a PicoTiterPlate: green circles represent pass filter wells, red circles are beads that contain the key pass (common adaptor) but the quality control considers that this reads are not good, blue and orange are control beads and purple represents empty beads, no common adaptor has been ligated. Chart D shows all the reads from a GS-FLX run and their distribution by length (bp).

BT.7

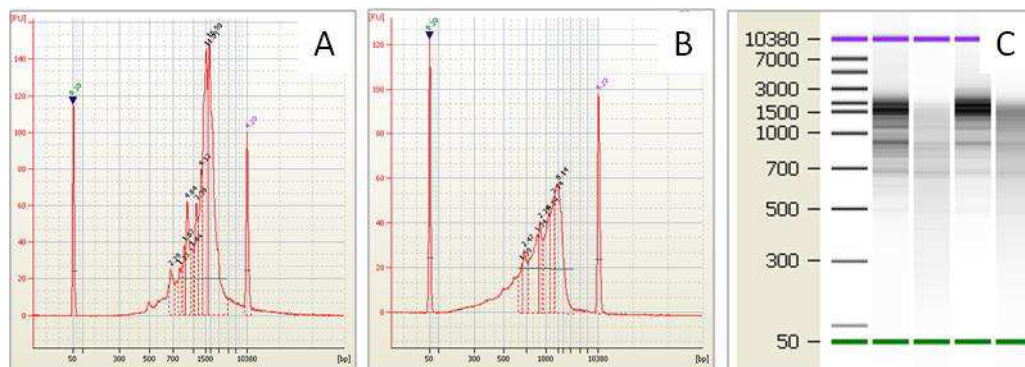
### 3. Examples of applications

“Next-generation” DNA sequencing is accelerating biological and biomedical research, by enabling the comprehensive analysis of whole genomes, transcriptomes, and interactomes to become inexpensive, routine and wide spread [17]. Important applications include novel whole genomes and transcriptomes sequencing, full-genome resequencing or more targeted discovery of mutations or polymorphisms, mapping structural rearrangements, DNA-protein interactions through chromatin immunoprecipitation sequencing, gene expression, discovering noncoding RNAs, ancient genomes, metagenomics, and other applications that will appear over the next few years.

#### 3.1. Novel whole genomes and transcriptomes sequencing

When we talk about genomic revolution, this is one of the applications that better shows the concept. During the last decade only researchers interested in human, rat, mouse and a few more had the genomic tools to develop their research since a reference genome and kits were available. NGS brings the opportunity to sequence what you want because no sequence knowledge is needed and it is relatively inexpensive and fast. For example, characterizing transcripts through sequences rather than through hybridization to a chip is advantageous in many ways, most importantly, because the sequencing approach does not require the knowledge of the genome sequence as a prerequisite and new transcripts can be described. From our experience in the Genomic Unit of the CCiTUB, a good strategy for a *de novo* transcriptome project consists in normalize the cDNA to generate full-length-enriched double stranded (ds) cDNA with equalized concentrations of different transcripts [18]. The cDNA normalization enhances the potential of the GS-FLX sequencing technology, since the equalized cDNA library can be sequenced without having an enormous

amount of repeated transcripts and obtaining a wide range of interested transcripts represented (Figure 3). Once this information is assembled and annotated, new genomic tools can easily be created (custom arrays, another NGS runs, Real-time PCR (see RT-PCR technology “Genomics: More than genes” in this issue)) to study gene expression under different conditions.



**Figure 3:** cDNA normalization results run in a Agilent 2100 Bioanalyzer at the Genomic Unit, CCIiTUB. Panel A shows ds cDNA not normalized, panel B shows the same sample normalized and the panel C shows a gel where the most predominant transcripts have partially been enzymatic digested obtaining an equalized concentrations of all different transcripts.

BT.7

### 3.2. Microsatellite discovery

Microsatellite discovery of non-model organism is another example where NGS are helping to open new areas of biological inquiry. Microsatellites are short segments of DNA that have a repeated sequence such as CACACACA, and they tend to occur in non-coding DNA. Over time, as animals in a population breed, they will recombine their microsatellites during sexual reproduction and the population will maintain a variety of microsatellites that is characteristic for that population and distinct from other populations which do not interbreed. NGS technologies are drastically reducing the time previously used by scientists to discover new markers to use in their research. Experiments performed in our lab showed that running 1/6 of a PicoTiterPlate in the GS-FLX system was able to generate over one hundred potential perfect microsatellites (di-, tri-, and tetranucleotides) with at least ten repeats ready to use by automated capillary electrophoresis (see Automated capillary electrophoresis “Genomics: More than genes” in this issue).

### 3.3. Targeted discovery of mutations or polymorphisms

For genomic resequencing, that is sequencing variation discovery in individuals of species for which a reference genome is available, a specific region of the genome across more individuals is frequently used in contrast to study the whole genome in fewer individuals. Some companies (NimbleGen, Qiagen, Fluidigm, RainDance) are developing different tools to target capture regions of the DNA. These strategies permit sequence specific genome regions to which a disease phenotype has been mapped, or exons of specific candidate genes belonging to disease, or the full complement of protein-coding DNA sequences.

## Acknowledgements

I would like to thank Agustin Arasan from Genomics Unit and Dr. Sergi Beltran from Bioinformatics Unit for their support in the optimization of the NGS technology at the CCIiTUB, Dr. Josep Planas to promote the development of the experimental design for *de novo* Transcriptome sequencing projects, and Dr. Creu Palacín, Dr. Rocío Pérez and Dr. Marta Pascual for the opportunity to participate in their Microsatellite discovering projects.

## References

- [1] Sanger F et al 1977 Proc. Natl. Acad. Sci. U.S.A. **74** 5463.
- [2] Maxam & Gilbert W 1977 Proc. Natl. Acad. Sci USA **74** 560
- [3] International Human Genome Sequencing Consortium. 2004. Nature **409**, 860
- [4] Margulies M, *et al.* 2005 Nature **437** 376.
- [5] Jonhsons DS et al 2007 Science **316** 5830
- [6] Picardi E 2010 Nucleic Acid Research **38** 14
- [7] Harris, T.D *et al.* 2008 Science **320**, 106.
- [8] Eid Jet al. 2009 Science **323**, 133
- [9] Metzker M. 2009 Nature reviews **27** 150
- [10] Adessi, C et al. 2000 Nucleic Acid Res. **28**, e87
- [11] Fedurco, M et al. 2006 Nucleic Acid Res. **34**, e22
- [12] Brenner, S et al. 2000 Nat. Biotechnol. **18**, 630
- [13] Mitra RD & Church 2003. Anal Biochem **320** 55
- [14] Mardis ER 2008. Annu Rev Genomics Hum Genet **9** 387
- [15] Metzker M. 2010 Nat Rev Genet **11** 31
- [16] Schuster SC 2008 Nature Methods **5** 16.
- [17] Shendure J & Hanlee J 2008 Nature biotechnology **26**, 1135
- [18] Shagin DA et al, 2002. Genome Res **12** 1935