Journal of Agriculture and Rural Research

NEXT-GENERATION HIGH-THROUGHPUT SEQUENCING TECHNOLOGIES: CONCEPT AND APPLICATIONS

M. S. Noman[1*], M. Rashid[1] and T. A. Khan[2]
[1] College of Plant Protection, China Agricultural University, Beijing, China.
[2] College of Agronomy and Biotechnology, China Agricultural University, Beijing, China
*Email: shibly.ent@gmail.com

Received: 14th April, 2017
Accepted: 25th May 2017
Published: 29th May, 2017

ABSTRACT

The development of DNA sequencing more than 35 years ago has profoundly impacted biological research. In the last few years, remarkable technological innovations have emerged that allow the direct and cost-effective sequencing of complex samples at unprecedented scale and speed. These next-generation technologies make it feasible to sequence not only static genomes, but also entire transcriptomes expressed under different conditions. These and other powerful applications of next-generation sequencing are rapidly revolutionizing the way genomic studies are carried out. In this article, the steps and applications of Next generation sequencing is discussed

Keywords: Agriculture, Application, Concept, DNA Sequencing

INTRODUCTION

Next-generation sequencing (NGS) refers to the deep, high-throughput, in-parallel DNA sequencing technologies developed a few decades after the Sanger DNA sequencing method first emerged in 1977 and then dominated for three decades (Sanger et al., 1977; Mardis, 2008) The NGS technologies are different from the Sanger method in that they provide massively parallel analysis, extremely high-throughput from multiple samples at much reduced cost [Mardis, 2011]. Millions to billions of DNA nucleotides can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that were used with Sanger sequencing (Metzker, 2010).The second-generation sequencing methods are characterized by the need to prepare amplified sequencing libraries before undertaking sequencing of the amplified DNA clones, whereas third-generation single molecular sequencing can be done without the need for creating the time-consuming and costly amplification libraries (Thompson and Milos, 2011). The parallelization of a high number of sequencing reactions by NGS was achieved by the miniaturization of sequencing reactions and, in some cases, the development of microfluidics and improved detection systems (Margulies et al., 2005). The time needed to generate the gigabase (Gb)-sized sequences by NGS was reduced from many years to only a few days or hours, with an accompanying massive price reduction.

The impact of NGS technology is indeed egalitarian in that it allows both small and large research groups the possibility to provide answers and solutions to many different problems and questions in the fields of genetics and biology, including those in medicine, agriculture, forensic science, virology, microbiology, and marine and plant biology.

Sanger sequencing and Next-generation sequencing

The principle behind Next Generation Sequencing (NGS) is similar to that of Sanger sequencing, which relies on capillary electrophoresis. The genomic strand is fragmented, and the bases in each fragment are identified by emitted signals when the fragments are ligated against a template strand.

The Sanger method required separate steps for sequencing, separation (by electrophoresis) and detection, which made it difficult to automate the sample preparation and it was limited in throughput, scalability and resolution. The NGS method uses array-based sequencing which combines the techniques developed in Sanger sequencing to process millions of reactions in parallel, resulting in very high speed and throughput at a reduced cost. The genome sequencing projects that took many years with Sanger methods can now be completed in hours with NGS, although with shorter read lengths (the number of bases that are sequenced at a time) and less accuracy (http://www.atdbio.com/content/58/Next-generation sequencing).

Steps of NGS:

Next generation methods of DNA sequencing have three general steps:

- Library preparation: libraries are created using random fragmentation of DNA, followed by ligation with custom linkers
- Amplification: the library is amplified using clonal amplification methods and PCR
- Sequencing: DNA is sequenced using one of several different approaches

a. Library Preparation

Firstly, DNA is fragmented either enzymatically or by sonication (excitation using ultrasound) to create smaller strands. Adaptors (short, double-stranded pieces of synthetic DNA) are then ligated to these fragments with the help of DNA ligase, an enzyme that joins DNA strands. The adaptors enable the sequence to become bound to a complementary counterpart.

Adaptors are synthesized so that one end is 'sticky' whilst the other is 'blunt' (non-cohesive) with the view to joining the blunt end to the blunt ended DNA. This could lead to the potential problem of base pairing between molecules and therefore dimer formation. To prevent this, the chemical structure of DNA is utilised, since ligation takes place between the 3′-OH and 5′-P ends. By removing the phosphate from the sticky end of the adaptor and therefore creating a 5′-OH end instead, the DNA ligase is unable to form a bridge between the two termini (Figure 1).

In order for sequencing to be successful, the library fragments need to be spatially clustered in PCR colonies or 'polonies' as they are conventionally known, which consist of many copies of a particular library fragment. Since these polonies are attached in a planar fashion, the features of the array can be manipulated enzymatically in parallel. This method of library construction is much faster than the previous labour intensive procedure of colony picking and E. coli cloning used to isolate and amplify DNA for Sanger sequencing, however, this is at the expense of read length of the fragments (http://www.atdbio.com/content/58/Next-generation sequencing).

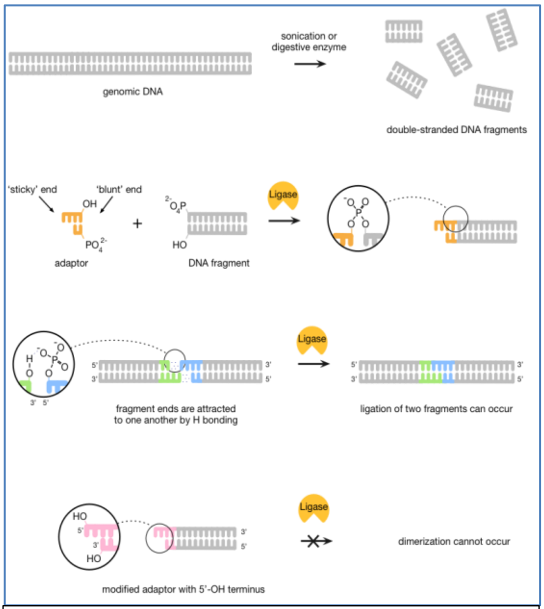
genomic DNA
sonication or digestive enzyme
double-stranded DNA fragments
'sticky' end
'blunt' end
OH
PO$_4^{2-}$
adaptor
+
$^{2-}$O$_4$P
HO
DNA fragment
Ligase
fragment ends are attracted to one another by H bonding
Ligase
ligation of two fragments can occur
HO
HO
modified adaptor with 5'-OH terminus
Ligase
dimerization cannot occur

Figure1.Library preparation of Next-generation sequencing

b. Amplification

Library amplification is required so that the received signal from the sequencer is strong enough to be detected accurately. With enzymatic amplification, phenomena such as 'biasing' and 'duplication' can occur leading to preferential amplification of certain library fragments. Instead, there are several types of amplification process which use PCR to create large numbers of DNA clusters.

1. Emulsion PCR

Emulsion oil, beads, PCR mix and the library DNA are mixed to form an emulsion which leads to the formation of micro wells (Figure 2).

In order for the sequencing process to be successful, each micro well should contain one bead with one strand of DNA (approximately 15% of micro wells are of this composition). The PCR then denatures the library fragment leading two separate strands, one of which (the reverse strand) anneals to the bead. The annealed DNA is amplified by polymerase starting from the bead towards the primer site. The original reverse strand then denatures and is released from the bead only to re-anneal to the bead to give two separate strands. These are both amplified to give two DNA strands attached to the bead. The process is then repeated over 30-60 cycles leading to clusters of DNA. This technique has been criticized for its time consuming nature, since it requires many steps (forming and breaking the emulsion, PCR amplification, enrichment etc) despite its extensive use in many of the NGS platforms. It is also relatively inefficient since only around two thirds of the emulsion micro reactors will actually contain one bead. Therefore an extra step is required to separate empty systems leading to more potential inaccuracies (http://www.atdbio.com/content/58/Next-generation sequencing).

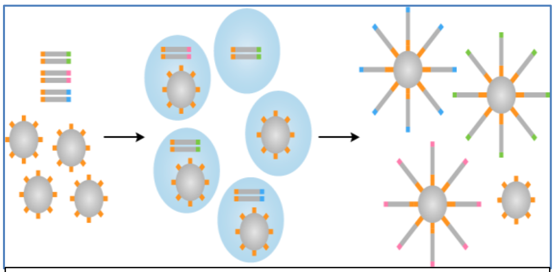

Figure 2. Emulsion PCR
Source: http://www.atdbio.com/content/58/Next-generation sequencing

2. Bridge PCR

The surface of the flow cell is densely coated with primers that are complementary to the primers attached to the DNA library fragments (Figure 3). The DNA is then attached to the surface of the cell at random where it is exposed to reagents for polymerase based extension. On addition of nucleotides and enzymes, the free ends of the single strands of DNA attach themselves to the surface of the cell via complementary primers, creating bridged structures. Enzymes then interact

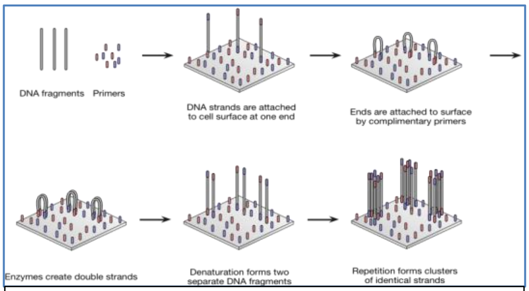
DNA fragments
Primers
DNA strands are attached to cell surface at one end
Ends are attached to surface by complimentary primers
Enzymes create double strands
Denaturation forms two separate DNA fragments
Repetition forms clusters of identical strands

Figure 3. Bridging PCR

Source: http://www.atdbio.com/content /58/Next-generation sequencing with the bridges to make them double stranded, so that when the denaturation occurs, two single stranded DNA fragments are attached to the surface in close proximity. Repetition of this process leads to clonal clusters of localised identical strands. In order to optimize cluster density, concentrations of reagents must be monitored very closely to avoid overcrowding.

c. Sequencing

Several competing methods of Next Generation Sequencing have been developed by different companies.

1. 454 Pyro sequencing

Pyro sequencing is based on the 'sequencing by synthesis' principle, where a complementary strand is synthesized in the presence of polymease enzyme (Figure 4). In contrast to using dideoxynucleotides to terminate chain amplification (as in Sanger sequencing), pyro sequencing instead detects the release of pyrophosphate when nucleotides are added to the DNA chain. It initially uses the emulsion PCR technique to construct the polonies required for sequencing and removes the complementary strand. Next, a ssDNA sequencing primer hybridizes to the end of the strand (primer-binding region), then the four different dNTPs are then sequentially made to flow in and out of the wells over the polonies. When the correct dNTP is enzymatically incorporated into the strand, it causes release of pyrophosphate. In the presence of ATP sulfurylase and adenosine, the pyrophosphate is converted into ATP. This ATP molecule is used for luciferase-catalysed conversion of luciferin to oxyluciferin, which produces light that can be detected with a camera. The relative intensity of light is proportional to the amount of base added (i.e. a peak of twice the intensity indicates two identical bases have been added in succession).

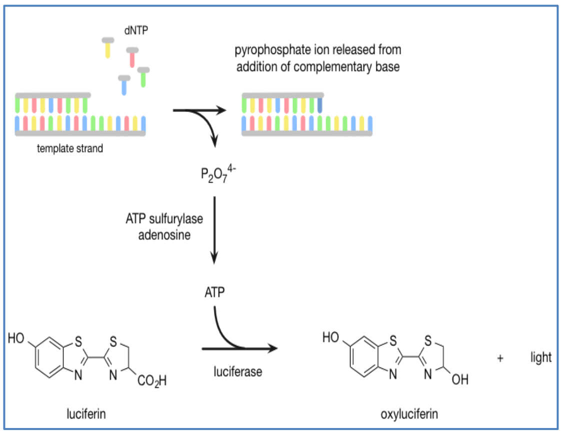
dNTP
pyrophosphate ion released from addition of complementary base
template strand
$P_2O_7^{4-}$
ATP sulfurylase adenosine
ATP
luciferase
HO
S
N
CO_2H
OH
+
light
luciferin
oxyluciferin

Figure 4. Pyro sequencing

Source: http://www.atdbio.com/content/58/Next-generation sequencing# Applications-of-Next-generation-sequencing

Pyro sequencing, developed by 454 Life Sciences, was one of the early successes of Next-generation sequencing; indeed, 454 Life Sciences produced the first commercially available Next-generation sequencer. However, the method was eclipsed by other technologies and, in 2013, new owners Roche announced the closure of 454 Life Sciences and the discontinuation of the 454 pyro sequencing platform (http://www.atdbio.com/content/58/Next-generation sequencing).

2. Ion torrent semiconductor sequencing

Ion torrent sequencing uses a "sequencing by synthesis" approach, in which a new DNA strand, complementary to the target strand, is synthesized one base at a time. A semiconductor chip detects the hydrogen ions produced during DNA polymerization (Figure 5).

Following polony formation using emulsion PCR, the DNA library fragment is flooded sequentially with each nucleoside triphosphate (dNTP), as in pyrosequencing. The dNTP is then incorporated into the new strand if complementary to the nucleotide on the target strand. Each time a nucleotide is successfully added, a hydrogen ion is released, and it detected by the sequencer's pH sensor. As in the pyrosequencing method, if more than one of the same nucleotide is added, the change in pH/signal intensity is correspondingly larger.

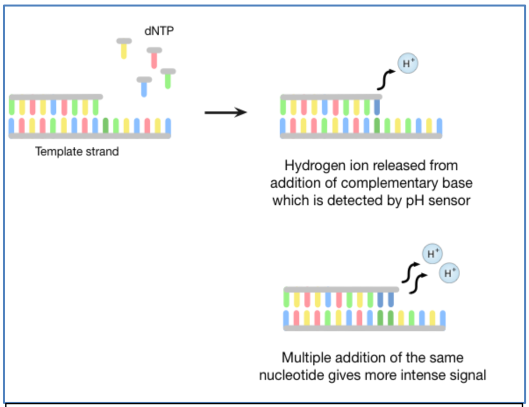
dNTP
H^+
Template strand
Hydrogen ion released from addition of complementary base which is detected by pH sensor
H^+
H^+
Multiple addition of the same nucleotide gives more intense signal

Figure 5. Ion Torrent semiconductor sequencing

Source: http://www.atdbio.com/content/58/Next-generation sequencing# Applications-of-Next-generation-sequencing

Ion torrent sequencing is the first commercial technique not to use fluorescence and camera scanning; it is therefore faster and cheaper than many of the other methods. Unfortunately, it can be difficult to enumerate the number of identical bases added consecutively. For example, it may be difficult to differentiate the pH change for a homorepeat of length 9 to one of length 10, making it difficult to decode repetitive sequences.

3. Sequencing by ligation (SOLiD)

SOLiD is an enzymatic method of sequencing that uses DNA ligase, an enzyme used widely in biotechnology for its ability to ligate double-stranded DNA strands (Figure 6). Emulsion PCR is used to immobilise/amplify a ssDNA primer-binding region (known as an adapter) which has been conjugated to the target sequence (i.e. the sequence that is to be sequenced) on a bead. These beads are then deposited onto a glass surface − a high density of beads can be achieved which which in turn, increases the throughput of the technique (http://www.atdbio.com/content/58/Next-generation sequencing).

Once bead deposition has occurred, a primer of length N is hybridized to the adapter, then the beads are exposed to a library of 8-mer probes which have different fluorescent dye at the 5' end and a hydroxyl group at the 3' end. Bases 1 and 2 are complementary to the nucleotides to be sequenced whilst bases 3-5 are degenerate and bases 6-8 are inosine bases. Only a complementary probe will hybridize to the target sequence, adjacent to the primer. DNA ligase is then uses to join the 8-mer probe to the primer. A phosphorothioate linkage between bases 5 and 6 allows the fluorescent dye to be cleaved from the fragment using silver ions. This cleavage allows fluorescence to be measured (four different fluorescent dyes are used, all of which have different emission spectra) and also generates a 5’-phosphate group which can ndergo further ligation. Once the first round of sequencing is completed, the extension product is melted off and then a second round of sequencing is perfomed with a primer of length N−1. Many rounds of sequencing using shorter primers each time (i.e. N−2, N−3 etc) and measuring the fluorescence ensures that the target is sequenced.

Due to the two-base sequencing method (since each base is effectively sequenced twice), the SOLiD technique is highly accurate (at 99.999% with a sixth primer, it is the most accurate of the second generation platforms) and also inexpensive. It can complete a single run in 7 days and in that time can produce 30 Gb of data. Unfortunately, its main disadvantage is that read lengths are short, making it unsuitable for many applications (http://www.atdbio.com /content/58/ Next-generation sequencing).

4. Reversible terminator sequencing (Illumina)

Reversible terminator sequencing differs from the traditional Sanger method in that, instead of terminating the primer extension irreversibly using dideoxynucleotide, modified nucleotides are used in reversible termination. Whilst many other techniques use emulsion PCR to amplify the DNA library fragments, reversible termination uses bridge PCR, improving the efficiency of this stage of the process.

Reversible terminators can be grouped into two categories: a. 3′-O-blocked reversible terminators and b. 3′-unblocked reversible terminators.

a. 3′-O-blocked reversible terminators

The mechanism uses a sequencing by synthesis approach, elongating the primer in a stepwise manner. Firstly, the sequencing primers and templates are fixed to a solid support. The support is exposed to each of the four DNA bases, which have a different fluorophore attached (to the nitrogenous base) in addition to a 3’-O-azidomethyl group (Figure 7).

Only the correct base anneals to the target and is subsequently ligated to the primer. The solid support is then imaged and nucleotides that have not been incorporated are washed away and the fluorescent branch is cleaved using TCEP (tris(2-carboxyethyl)phosphine). TCEP also removes the 3’-O-azidomethyl group, regenerating 3’-OH, and the cycle can be repeated (Figure 8).

b. 3′-unblocked reversible terminators

The reversible termination group of 3′-unblocked reversible terminators is linked to both the base and the fluorescence group, which now acts as part of the termination group as well as a reporter. This method differs

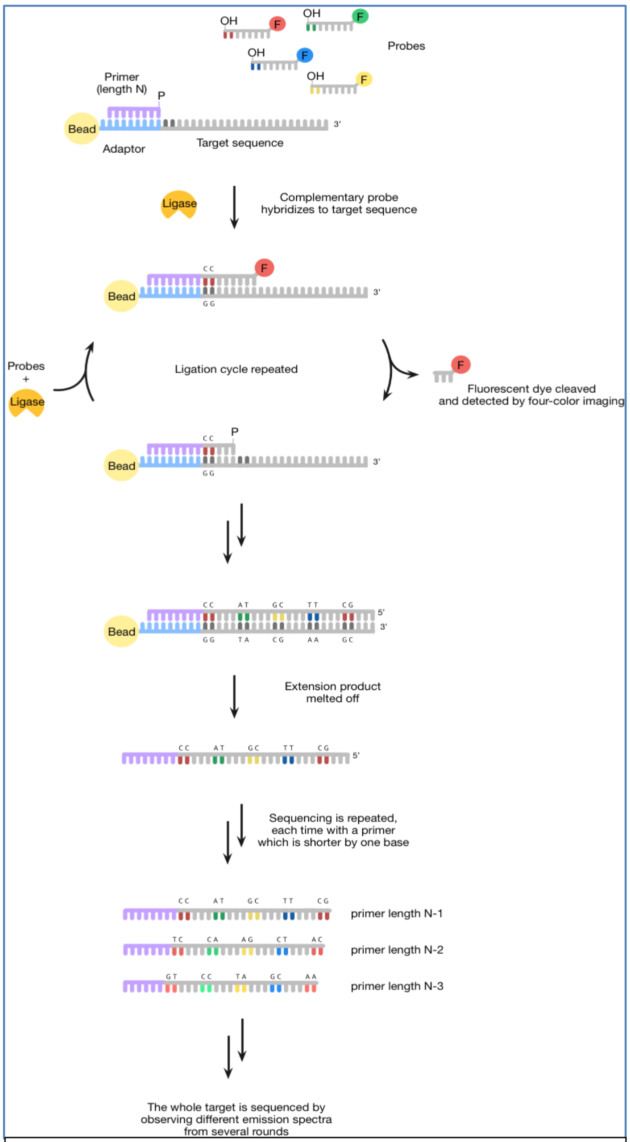
OH
F
OH
F
Probes
OH
F
OH
F
Primer (length N)
P
Bead
3'
Adaptor
Target sequence
Ligase
Complementary probe hybridizes to target sequence
CC
F
Bead
3'
GG
Probes + Ligase
Ligation cycle repeated
F
Fluorescent dye cleaved and detected by four-color imaging
CC
P
Bead
3'
GG
CC AT GC TT CG
Bead
5'
3'
GG TA CG AA GC
Extension product melted off
CC AT GC TT CG
5'
Sequencing is repeated, each time with a primer which is shorter by one base
CC AT GC TT CG
primer length N-1
TC CA AG CT AC
primer length N-2
GT CC TA GC AA
primer length N-3
The whole target is sequenced by observing different emission spectra from several rounds

Figure 6. Sequencing by ligation (Source: http://www.atdbio.com/content/58/Next-generation sequencing# Applications-of-Next-generation-sequencing)

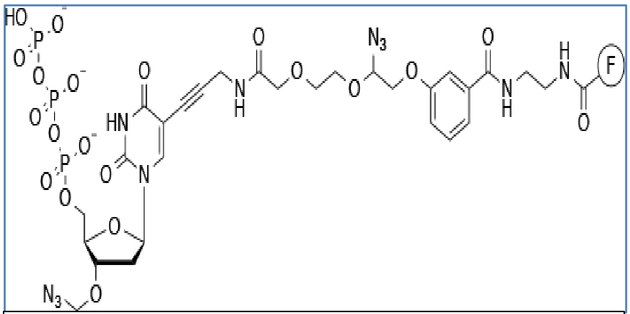

Figure7. Structure of fluorescently labelled dNTP used in Illumina sequencing

from the 3′-O-blocked reversible terminators method in three ways: firstly, the 3’-position is not blocked (i.e. the base has free 3’-OH); the fluorophore is the same for all four bases; and each modified base is flowed in sequentially rather than at the same time.

The main disadvantage of these techniques lies with their poor read length, which can be caused by one of two phenomena. In order to prevent incorporation of two nucleotides in a single step, a block is put in place, however in the event of no block addition due to a poor synthesis, strands can become out of phase creating noise which limits read length. Noise can also be created if the fluorophore is unsuccessfully attached or removed. These problems are prevalent in other sequencing methods and are the main limiting factors to read length.

This technique was pioneered by Illumina, with their HiSeq and MiSeq platforms. HiSeq is the cheapest of the second generation sequencers with a cost of $0.02 per million bases. It also has a high data output of 600 Gb per run which takes around 8 days to complete (http://www.atdbio.com/content/58/Next-generation sequencing).

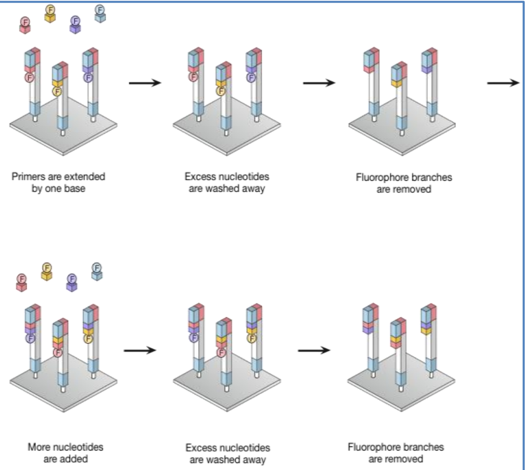
F
Primers are extended by one base
Excess nucleotides are washed away
Fluorophore branches are removed
More nucleotides are added
Excess nucleotides are washed away
Fluorophore branches are removed

Figure 8. Reversible terminator sequencing Source: http://www.atdbio.com/content/58/Next-generation sequencing# Applications-of-Next-

Current and prospective applications of next-generation sequencing technologies

In addition to the expected benefits of cost-effective DNA sequence information at revolutionary depth, scale and throughput, unexpected benefits have been derived from the new sequencing technologies: (1) the analysis of gene expression by transcriptomic profiling, and (2) the analysis of mechanisms behind the regulation of gene expression by epigenomic profiling. Altogether, novel applications in genome wide genetic variation, transcriptomic and epigenomic analyses position the next generation sequencing technologies as ground breaking integrative tools providing unprecedented insights into genome-wide functional genomics.

Fundamental advances in genetics and genomics, transcriptomics and epigenomics have repercussions in virtually all fields of biology, with downstream applications in medicine and nutrition, plant and animal breeding and agriculture biotechnology. Furthermore, the new sequencing technologies have opened the gates to a new world of understanding of microbial diversity and ecology, with great promise for innovative applications in agriculture, nutrition, alternative energy production and the environment.

A. Genome Level Applications

1. De novo sequencing with the next-generation technologies

De novo genome sequencing refers to the sequencing of genomes for which there is no prior sequence information. To decipher a genome without prior information, individual reads must be assembled based on sequence overlaps only.

For small and simple genomes, the new sequencing technologies provide a fast and economic alternative to Sanger sequencing: the GS FLX System in particular, which produces the longest reads of the new technologies, has successfully generated de novo sequences of whole bacterial and archeal genomes.

For larger and more complex genomes, however, de novo sequence assembly represents a major bottleneck for the new technologies. Though laborious and expensive, Sanger technology remains better equipped for de novo sequencing of complex genomes, owing to the production of long reads and the possibility for clone by-clone or hierarchical sequencing strategy. Nevertheless, the new sequencing technologies do bring a major breakthrough for unsequenced genomes of any size and complexity, with the possibility to generate cost-effective genome-wide sequence information for marker or gene discovery.

Applications:

DNA sequence is the foundation for genomics research, at the basis for rational breeding at the molecular level for improved yield, quality and sustainability of agricultural products. Furthermore, obtaining a whole genome sequence is a pre-requisite to access a treasure trove of functional data generated by the whole-genome sequence-based applications of the new sequencing technologies.

1) Direct applications to whole-genome sequencing of simple genomes:

- Sequencing of “exotic” organisms that may use novel metabolic pathways of interest, e.g., for adaptation to poor soils (Alcaraz et al., 2008) or for converting biomass into energy (Hongoh et al., 2008).
- Sequencing of key uncultured bacteria/archae to provide reference genomes for genetic diversity analyses and to facilitate metagenomic analyses (Hongoh et al., 2008).

2) Current applications to eukaryote genomes through the production of partially assembled yet informative sequences:

- De novo transcriptome sequencing provides a first survey of exon sequences from uncharacterized genomes, instrumental for gene discovery, annotation, and SNP (single nucleotide polymorphism) discovery (Novaes et al., 2008).
- Direct sequencing of genomic fragments enable SNP discovery irrespective of gene expression in non-model organisms (Bekal et al., 2008)

3) Foreseen applications, as eukaryote genomes become amenable to direct de novo sequencing by improved technologies:

- ○ Pursue the sequencing of model organisms as the basis for basic research in structural and functional genomics more cost-effectively.
- ○ Sequencing of major farm animals and crop plants to harness all the potential of molecular breeding
- ○ Sequencing of key evolutionary nodes in the phylogenetic tree for research in plant and animal evolution and domestication

2. Facilitated sequencing of related genomes and resequencing

Whole-genome sequencing is simplified when the genome of a closely related species is known. The related genome sequence is used as a reference genome or scaffold onto which short sequences can be aligned, greatly facilitating the new genome assembly. When the reference genome is from the same species as the genome to be sequenced, sequencing becomes a resequencing exercise, where the key to the variant genome reconstruction strictly lies in the correct mapping of the new reads to the reference genome.

Applications:

1) Exploitation of related genetic resources

Facilitated sequencing of related species enables inter-specific comparative genomics for evolutionary, phylogenetic and functional analysis, and helps determine genetic targets for breeding at the molecular level, e.g.:

- ○ *Vitis vinifera* grapevine genome, sequenced with Sanger technology, is used as a reference scaffold for high-throughput sequencing of the related species *Vitis riparia,* of interest as a source of resistance to diseases (e.g., phylloxera).
- ○ Similarly, the genome sequence of cultivated tomato *Solanum lycopersicum* (previously *Lycopersicon esculentum*) could enable cost-effective sequencing of wild relative *S. pennellii* (previously *L. pennellii*), which would help reveal the molecular nature of complex agronomic traits exposed in the introgression lines of S. pennellii chromosomal segments in *S. lycopersicum*.

2) Genetic diversity, ancient DNA and evolutionary studies:

- ○ resequencing multiple strains of model organisms to assess genetic diversity (Illumina proof-of-concept paper) (Hillier et al., 2008)
- ○ Applications of ancient DNA analysis: high-throughput, cost-effective technologies allowing deep sequencing is essential for ancient DNA analysis considering it is often degraded and in minute concentration in specimen samples. Applications include:
- - resequencing ancient mitochondrial genomes, to reveal historical population dynamics (Gilbert et al., 2008)
- - improved phylogenetic understanding, relationship between extinct and modern species analysis of domestication process
- - impact analysis of past climatic changes on soil community and plant and animal distribution
- - the potential for “reactivation” of ancient genes by transgenesis

3) Genome-wide discovery of genetic variation as a pre-requisite for genotyping applications (e.g., population structure analysis, genotype-phenotype association studies, marker-assisted breeding, personal medicine, etc., developed in the dedicated Genotyping worksheet):

- ❖ Foreseen application: resequencing entire core collections of plant and animal genetic resources to uncover most germplasm genetic diversity for subsequent genotyping and breeding applications.
- ❖ Current applications to livestock and crops involve complexity-reduction strategies to reduce the size and cost of experiments:
- ○ deep resequencing of bovine reduced-representation libraries identifies large numbers of genome-wide SNPs in target populations (Van Tassell et al., 2008)
- ○ resequencing expressed genes (ESTs) for cost-effective discovery of SNPs between two lines representing major heterotic groups of maize (Barbazuk et al., 2007)
- ○ Application to catalog human genetic variation is ongoing with the 1000 genomes project (www.1000genomes.org).

4) Discovery of genetic variants associated with a phenotype, e.g.:

- ○ resequencing strains of Mycobacterium tuberculosis sheds light on antibiotic resistance (Andries et al., 2005)
- ○ strain-to-reference comparison identifies markers for reactive biosecurity applications (La Scola et al., 2008)
- ○ Whole-genome mutational profiling after mutagenesis breeding, e.g., on Pichia stipitis bred for improved xylose-to-ethanol conversion (Smith et al., 2008)

5) Detection of rare somatic mutations by ultra-deep resequencing

Successfully applied to the discovery of somatic mutations during tumor development (Campbell et al., 2008), applications are foreseen to assess somaclonal mutations during clonal propagation of plants and trees.

B. Transcriptome Level Applications

Next-generation sequencing technologies can do more than providing raw sequence information: applied at the transcriptome level, they can replace microarray-based strategies and provide an open, digital platform for genome-wide analysis of gene expression, without relying on previous annotation data. Applied to a non-sequenced genome, transcriptome sequencing with long reads provides a first access to gene diversity. Applied to a sequenced genome, higher throughput short read sequencing provides deep quantitative analysis of gene expression on a genome scale.

1. De novo transcriptome sequencing for broad gene and marker discovery

Sequencing normalised full-length complementary DNA generated from pools of RNA extracted from diverse tissues, conditions and genotypes can provide a first survey of genetic diversity and a large set of genetic markers for species with no prior sequence information. This approach requires long reads to facilitate contig assembly and reconstitute transcripts sequence. Although generating shorter reads than Sanger, GS-FLX was shown to uncover more genetic diversity than ABI 3730.

Applications:

1) Transcript profiling without prior sequence knowledge

2) Gene sequence analysis of plants or animals whose genome complexity (large size, numerous repeated elements, polyploidy) excludes a genome project, as is the case for pea.

3) Production of new expressed sequences for gene discovery, functional analysis, annotation and SNP discovery (Novaes et al., 2008).

2. Quantitative gene expression profiling

The so-called RNA-seq procedure enables deep quantitative analysis of gene expression of any sequenced genomes (Graveley, 2008). Here, the sequence is not the primary interest. In fact, knowing the sequence of the genome to be analysed is a prerequisite for RNA-seq. Sequence reads should be just long enough to be mapped on a reference genome. The interest is where the sequence maps, and how many reads map there, which provides hypothesis-free information on transcriptional units, splicing and expression levels. This approach is made possible due to the new technologies capacity to map and count reads following sequencing at great depth.

Applications:

1) Digitally measure the presence and prevalence of transcripts from known and previously unknown genes
Discovery of novel transcriptional units and alternative splicing

3) Possibility to distinguish sense from anti-sense transcripts

4) Precise measure of gene expression level and distinction between members of gene families

5) Deep sequencing to identify low-abundance transcripts

6) Cost-effective technique to profile transcriptomes in different mutants, different tissues, under different conditions (Lister et al., 2008; Wilhelm et al., 2008)

3. Deep sequencing of small RNAs

Specific to eukaryotes, small non-coding RNAs are key regulators of a number of biological processes including development, stress responses and genome stability. Small RNAs also play an important role in transgene expression. There are different types of small RNAs, mostly belonging to two major groups: (1) the microRNAs (miRNAs), generated from broadly conserved MIR genes, which act as negative regulators of target genes by degrading, or inhibiting the translation of, complementary target mRNAs, and (2) the short or small interfering RNAs (siRNAs), generated from DNA repeats, transposons or incorrectly processed RNA transcripts, which are involved in gene silencing by guiding novel epigenetic modifications, or through mRNA cleavage.

Small RNAs can be specifically analysed in a dedicated smRNA-seq approach, applied to a fraction of total RNA, usually within a 15-30 nt interval. Specific bioinformatics pipelines enable to distinguish between types of small RNAs.

Applications:

1) Discovery of novel, mostly low-abundance, less-conserved miRNAs

o organ-specific (e.g., rice grain, tomato fruit) (Moxon et al., 2008; Zhu et al., 2008)

o lineage and/or species-specific (e.g., avian/chicken) (Glazov et al., 2008)

o abiotic or biotic stress-specific

2) Investigation of the role of small RNAs in development patterning, in lineage and/ or species-specific pathways and functions, in abiotic and biotic stress responses

3) Subsequent exploitation of miRNAs and RNA interference to modulate the expression of target genes of interest

C. Epigenome Level Applications

Next-generation sequencing reveals genome-wide profiles of gene expression, but can it help understand how gene expression is regulated? The regulation of gene expression is partly controlled at the epigenetic level by modifications that affect gene expression without affecting the DNA sequence. Epigenetic modifications typically include DNA methylation, post-translational modification of histone proteins, and variations in nucleosome positioning. Unexpectedly, the next-generation sequencing technologies can not only detect but precisely map and quantify these modifications.

Sequencing then enters new realms of investigation that of (1) the fundamental mechanisms of regulation of gene expression by epigenetic modifications, (2) cell differentiation and specialisation during normal development, and (3) the specific modulation of gene expression upon environmental triggers. These issues can now be addressed on a genome scale, across tissues, across treatments, developmental stages and generations, for a better understanding of how these modifications are acquired, orchestrated and inherited, for what consequences, and whether the ultimate code for these non-genetic modifications is not controlled in the DNA sequence after all.

1. DNA Methylation profiling

Cytosine methylation in specific sequence contexts causes stable and heritable gene silencing. DNA methylation can be analysed at single-base resolution by sequencing bisulphite-treated DNA, a procedure termed MethylC-seq or BS-seq, depending on the publications. Bisulphite treatment converts unmethylated cytosines to uracils, while methylated cytosines remain unchanged, enabling to deduce the status of each cytosine by straight sequencing. Other techniques assessing cytosine methylation exist that are currently cheaper but less exhaustive (using methylation-sensitive endonucleases), or less precise (using methyl-C immunoprecipitation).

Applications:

1) Reveal new methylation sites at single-base resolution, inform on global methylation patterning or specific methylated promoters (Cokus et al., 2008)

2) Explore different pathways and regulation of methylation in different sequence context by profiling DNA methylation in different DNA methylation mutants (Cokus et al., 2008)

3) Integrate genome-wide DNA methylation, small RNA, and mRNA profiles to analyse the global interplay of epigenetic modifications, RNA interference and transcription (Lister et al., 2008).

2. Protein-DNA interaction profiling

The so-called ChIP-seq procedure enables to identify and quantify in vivo protein-DN A interactions on a genome scale. ChIP-seq combines high-throughput sequencing with chromatin immunoprecipitation (ChIP): sequencing DNA fragments immunoprecipitated with a DNA-binding protein of interest enables high-resolution mapping of binding sites to any sequenced genome (Figure 9).

DNA-binding proteins of interest include transcription factors and histone proteins. Histones may undergo a large range of post-translational modifications, which are proposed to function combinatorially or sequentially to regulate downstream functions, such as marking enhancer elements or guiding DNA methylation.

Applications:

1) Genome-wide identification of DNA targets of transcription factors, for different cell types and physiological conditions (Johnson et al., 2007)

2) Genome-wide profiling of binding sites of modified histones in different tissues, different conditions, or during differentiation of pluripotent stem cells to understand the epigenetic control of cell specialisation (Mikkelsen et al., 2007)

3. Nucleosome positioning

The position of nucleosomes directly influences gene regulation by controlling access of transcription factors and transcription machinery to the DNA sequence. Various strategies involving sequencing mononucleosomal DNA with all three nextgeneration technologies enable to map the positions of nucleosomes at high resolution throughout the genome, giving unprecedented data sets for inferring positioning rules in relation to DNA sequence (Albert et al., 2007).

Application: to acquire a better understanding of how genes are regulated by nucleosome positioning, and how nucleosome positioning is controlled.

Conclusion

Next-generation sequencing (NGS) technologies using DNA, RNA, or methylation sequencing have impacted enormously on the life sciences. NGS is the choice for large-scale genomic and transcriptomic sequencing because of the high-throughput production and outputs of sequencing data in the gigabase range per instrument run and the lower cost compared to the traditional Sanger first-generation sequencing method. NGS today is more than ever about how different organisms use genetic information and molecular biology to survive and reproduce with and without mutations, disease, and diversity within their population networks and changing environments. Next-Generation Sequencing has changed the way we carry out molecular biology and genomic studies. It has allowed us to sequence and annotate genomes at a much faster rate. It has allowed us to study variation, expression and DNA binding at a genome-wide level.

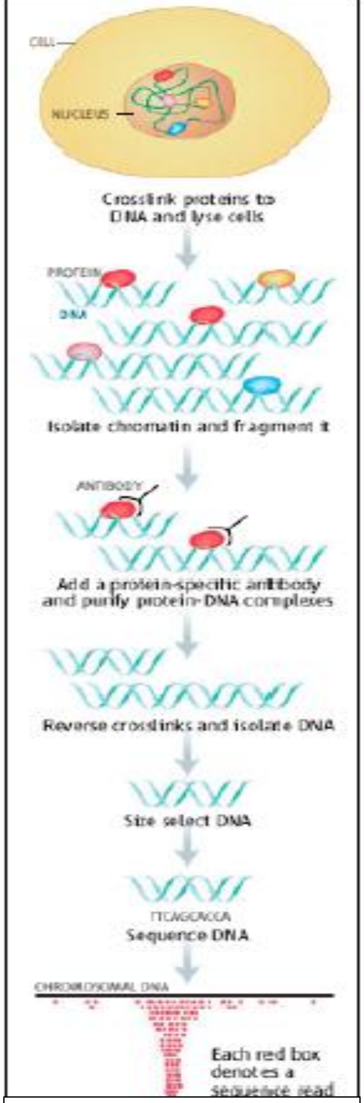
CELL
NUCLEUS
Crosslink proteins to DNA and lyse cells
PROTEIN
DNA
Isolate chromatin and fragment it
ANTIBODY
Add a protein-specific antibody and purify protein-DNA complexes
Reverse crosslinks and isolate DNA
Size select DNA
TTCAGCACCA
Sequence DNA
CHROMOSOMAL DNA
Each red box denotes a sequence read

Figure 9. Protein-DNA interaction profiling (Johnson et al., 2007)

References

Ahmadian, A, Svahn H and Parallel M 2011.Sequencing Platforms using Lab on a Chip Technologies. *Lab Chip*, 11: 2653 − 2655

Alcaraz LD, Olmedo G, Bonilla G, Cerritos R, Hernandez G, Cruz A, Ramirez E, Putonti C, Jimenez B and Martinez E 2008. The genome of Bacillus coahuilensis reveals adaptations essential for survival in the relic of an ancient marine environment. *Proc Natl Acad Sci U S A* 105: 5803-5808.

Andries K, Verhasselt P, Guillemont J, Gohlmann H W, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E, Williams P, de Chaffoy D, Huitric E, Hoffner S, Cambau, E, Truffot-Pernot C, Lounis N and

Jarlier V 2005. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. *Sci* 307: 223-227.

Balasubramanian S 2011. Decoding Genomes at High Speed: Implications for Science and Medicine. *Angew. Chem Int. Ed*, 50: 12406-12410

Balasubramanian S 2011. Sequencing Nucleic Acids: from Chemistry to Medicine. *Chem. Commun*, 47:7281 − 7286.

Barbazuk WB, Emrich SJ, Chen HD, Li L and Schnable PS 2007. SNP discovery via 454 transcriptome sequencing. *Plant J,* 51: 910-918.

Bekal S, Craig JP, Hudson ME, Niblack TL, Domier LL and Lambert KN 2008. Genomic DNA sequence comparison between two inbred soybean cyst18 nematode biotypes facilitated by massively parallel 454 micro-bead sequencing. *Mol Genet Genomics* 279:535-543.

Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA and Stratton MR 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A* 105: 13081-13086.

Chen F, Dong M, Ge M, Zhu L, Ren L, Liu G, Mu R 2013. The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Gen. Pro. Bio*. 11: 34-40.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M and Jacobsen SE 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452: 215-219.

Gilbert MT, Kivisild T, Gronnow B, Andersen PK, Metspalu E, Reidla M, Tamm E, Axelsson E, Gotherstrom A and Campos PF 2008. Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. *Sci,* 320: 1787-1789.

Graveley BR 2008. Molecular biology: power sequencing. *Nature* 453: 1197-1198.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M and Huang W 2008. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 5: 183-188.

Hongoh Y, Sharma VK, Prakash T, Noda S, Taylor TD, Kudo T, Sakaki Y, Toyoda A, Hattori M and Ohkuma M 2008. Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci U S A,* 105: 5555-5560.

Johnson DS, Mortazavi A, Myers RM, and Wold B 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Sci* 316: 1497-1502.

La Scola B, Elkarkouri K, Li W, Wahab T, Fournous G, Rolain JM, Biswas S, Drancourt M, Robert C, Audic S, Lofdahl and Raoult D 2008. Rapid comparative genomic analysis for clinical microbiology: The Francisella tularensis paradigm. *Genome Res* 18: 742-750.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH and Ecker JR 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133: 523-536.

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

Mardis ER 2011. A decade’s perspective on DNA sequencing technology. *Nature*, 470:198–203. DOI: 10.1038/nature09796

Margulies M, Egholm M, Altman WE 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–80. PMID: 16056220

Metzker ML 2010. Sequencing technologies — the next generation. *Nat Rev Genet*, 11:31–46. DOI: 10.1038/nrg2626

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553-560

Moxon S, Jing R, Szittya G, Schwach F, Rusholme Pilcher RL, Moulton V and Dalmay T 2008. Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Res*., in press.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr, Grattapaglia D, Sederoff RR and Kirst M 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome, *BMC Genomics* 9: 312.

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

Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.,* in press.

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

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC and Sonstegard TS 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods,* 5: 247-252.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, and Bahler J 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature,* 453: 1239-1243.

Zhu Q H, Spriggs A, Matthew L, Fan L, Kennedy G, Gubler F and Helliwell C 2008. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res,* 18: 1456-1465.