

Computational microbial genomics

Giacomo Fantoni

telegram: @GiacomoFantoni

Github: <https://github.com/giacThePhantom/computationl-microbial-genomics>

April 5, 2022

Contents

1	<i>Escherichia Coli</i> general informations	2
1.1	<i>E. coli</i> genomics	2
1.1.1	<i>E. coli</i> long-term evolution experiment	2
1.1.2	<i>E. coli</i> strains	2
1.1.3	PanPhlAn - strain detection and characterization	3
1.1.4	Genomes of <i>E. coli</i>	3
2	NGS principles (second gen. sequencing) - From Sanger to third gen sequencing	5
2.1	History of Sequencing	5
2.1.1	Progresses of sequencing	5
2.1.2	The Chain Terminators	6
2.1.3	Sanger method: the first one	6
2.2	Development of Sequencing Machines	8
2.3	Next Generation Sequencing NGS	10
2.3.1	Fragments/Library preparation	11
2.3.2	Clonal amplification and ILLUMINA sequencing procedure	11
2.3.3	Pacific Bioscience (PacBio)	15
2.3.4	Nanopore sequencing	15
2.4	Sequence mapping	16
2.4.1	Coverage	16
2.4.2	Smit-Waterman algorithm	16
2.4.3	Needleman-Wunsch algorithm	16
2.5	Blast	16
2.5.1	Scoring matrices	17
2.6	Spaced seed alignment	17
2.7	Burrows-Wheeler transform	17
3	Sequencing Data	18
3.0.1	FASTQ format	19
3.0.2	Quality control	19
4	<i>Staphylococcus aureus</i>	20

Chapter 1

Escherichia Coli general informations

1.1 *E. coli* genomics

Escherichia Coli is a Gram-negative, facultative anaerobic, rodshaped, coliform bacterium, it pertains to the phylum of proteobacteria and to the family of Enterobacteriaceae. It can be grown easily and inexpensively. It has got genome with a length between 4.5 - 4.7 M bases, including about 4000-5000 genes, and about seven ribosomal RNA operons. Only the 38% of the genes of K-12 (one of the most studied bacterial strains of *E. coli*) were experimentally identified, overall 40-50% of the genes are to date without a known function. The original *E. coli* strain K-12 was obtained from a stool sample of a diphtheria patient in Palo Alto, CA in 1922.

1.1.1 *E. coli* long-term evolution experiment

The *E. coli* long-term evolution experiment led by Richard Lenski is one of the longest evolutionary experiments ever made (search "The Longest-Running Evolution Experiment"). The experiment started on 24th February 1988, and since that moment 12 populations of *E. coli* have been cultivated in the same environment. After each day (corresponding to the time of development of approximately 7 generations), a portion of bacteria from each flask was introduced in a new one, and let proliferating in it. Every 500 generation, it has been saved a sample of the bacteria of each flask, in order to track the evolutionary changes made. Today the experiment is on-going, and researchers reached approximately the 66000th generation. The study suggests a series of conclusions, to cite "long-term adaptation to a fixed environment can be characterized by a rich and dynamic set of population genetic processes, in stark contrast to the evolutionary desert expected near a fitness optimum" (Good et al 2017). In fact, despite of the fixed environment, some bacteria developed the capacity to aerobically grow on citrate, which is unusual in *E. coli* (around generation 31,000) and developed complex mutation patterns.

1.1.2 *E. coli* strains

E. coli could be found as commensal strains, pathogenic strains, or environmental strains. The pathogenic strains could pertain to these categories (which are not exclusive): enteropathogenic

1.1. E. COLI GENOMICS

(EPEC), enteroinvasive (EIEC), enterotoxigenic (ETEC), diffusely adherent (DAEC), adherent invasive (AIEC), shiga-toxin producing (STEC), enteroaggregative(EAEC), extraintestinal pathogenic (ExPEC). Resistances to antibiotics make even more difficult the process of categorization of *E. coli*. In 2011 in Germany, an outbreak of Stx-EAEC was responsible of the death of some people. An efficient counter-measure was found by sequencing the genome of those bacteria.

Shigella is *E. coli* with shiga toxin. It had been an issue for taxonomists.

Most of the genes are on plasmids, circular, additional to chromosome, and can be moved easily horizontally. Plasmids between different strains can be moved in enterobacteriaceae, this doesn't happen normally in other families. Some *E. coli* strains are even capable of causing tumors in humans: for example, colibactin-positive *E. coli* can cause colon and rectal cancer, by creating mutations which are responsible of the cancer onset.

several antigens can be used by taxonomists to categorize *E. coli* strains. In particular there are the O, H, K antigens, respectively related to the somatic, the flagella and the capsule. O antigens are 171, Ks are 80 and Hs are 56.

1.1.3 PanPhlAn - strain detection and characterization

Pangenome-based Phylogenomic Analysis (PanPhlAn) is a strain-level metagenomic profiling tool for identifying the gene composition and in-vivo transcriptional activity of individual strains in metagenomic samples. PanPhlAn's ability for strain-tracking and functional analysis of unknown pathogens makes it an efficient tool for culture-free infectious outbreak epidemiology and microbial population studies (PanPhlAn reference). This tool was for example used to study the strain responsible of an outbreak in Germany in 2011. This strain was a shiga-toxigenic Escherichia coli (STEC), and the study was conducted by Loman and colleagues in 2013. This method outlasted the traditional one.

sequencing means generally to sequence everything, it's normally difficult where to find it, although in *E. coli* is quite easy to understand the provenience. from all the world, populating diversity of *E. coli*, every time we sample we find points which are different from the reference genomes. points overlapping are people living together and share bacterias

1.1.4 Genomes of *E. coli*

In the genome of *E. coli* strains, it is possible to distinguish:

- **Core genome:** the set of all genes shared by all members of a bacterial species, it includes 1000 up to 3000 genes.
- **Accessory or dispensable genome:** the set of genes present in some but not all genomes within the same bacterial species. found on a single strain or in a subset of strains.
- **Pangenome:** core genome + accessory genome. set of all the genes foundable in the species strains. It is said to be "*closed*" when pangenome size tends to a maximum as number of genomes increases, instead it is "*open*" when pangenome keeps increasing as you add new genomes

Sequencing more organisms of the same species means to lower the amount of genes in the core genome and augment the number of those in the pan-genome (figure ??). Because of technical problems the probability of getting a gene and not forget it is different from 0, so why probably the sequencing of other genomes would lead technically to a plummet to 0 of the pangenome. With some mathematical formulations we can predict a more probable plateau (Rasko David 2008).

1.1. *E. COLI* GENOMICS

each *E. Coli* genome contains in a balanced way genes of the core genome and of the pan genome, for a total amount of genes correspondant to about 4700 genes (figure ??). Core genomes' genes are responsible of some basic cellular functionalities and utilities to survive environment, while instead elements of the pangenome are quite usually specific to the single strains, they are not always functionally well known.

ratios of the pan-genome and the core-genome are not equal in other organisms behave differently.

Chapter 2

NGS principles (second gen. sequencing) - From Sanger to third gen sequencing

Figure 2.1

- 1959 – First homogenous DNA purified
- 1970 – First discovery of type II restriction enzymes
- 1972 – First RNA gene sequence published (lac operon)
- 1975 – Sanger first publishes his plus/minus method of sequencing (unable to distinguish homopolymers)
- 1977 – Maxam & Gilbert publish their method (could distinguish homopolymers)
- 1977 – Sanger publishes Dideoxy sequencing method

NGS stands for ***Next Generation Sequencing***, and it represents the method of sequencing most used nowadays. Before that, a series of other discoveries were done, elenicated in the figure 2.1.

2.1 History of Sequencing

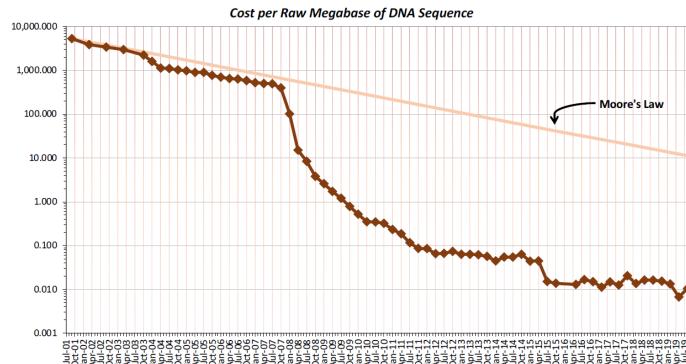
2.1.1 Progresses of sequencing

As the time passes, the cost of sequencing the DNA is diminishing. The rate of decrease actually is higher than the one predicted by the **Moore's Law**, "*Democratization of sequencing*" is seen nowadays, because of the costs lowering. After the Human genome project, several other animals and plants' genomes were sequenced.

The methods of sequencing actually can be grouped in three **groups**, that are:

2.1. HISTORY OF SEQUENCING

Figure 2.2



- **Chemical degradation** of DNA: it includes the method of Maxam-Gilbert
- **Sequencing by synthesis (“SBS”)** which is the most common approach and the first to be developed. It uses DNA polymerases in primer extension reactions Illumina, Pacific Biosciences, Ion Torren and 454
- **Ligation-based:** sequencing using short probes that hybridize to the template, the technologies pertaining to this class are SOLiD, Complete Genomics
- **Other:** Nanopores

2.1.2 The Chain Terminators

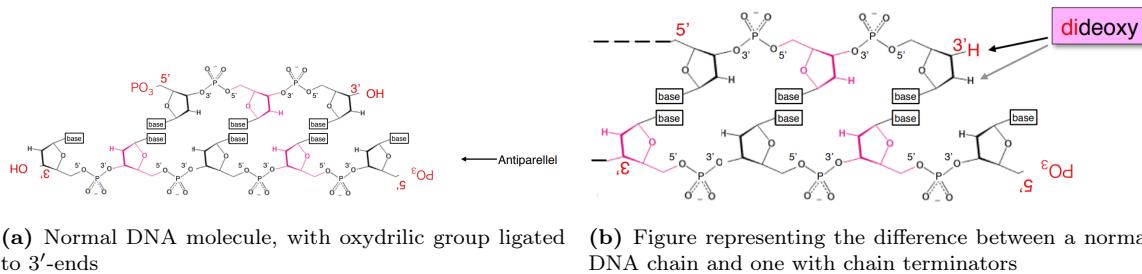


Figure 2.3: Normal DNA synthesys vs Chain terminators

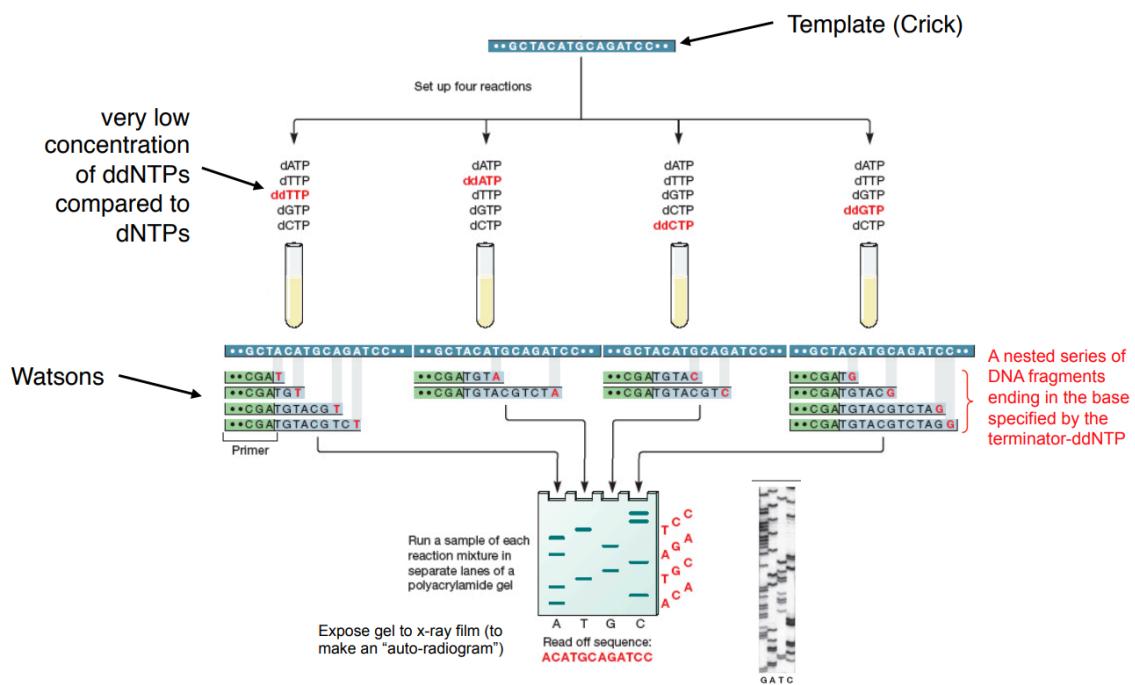
Normally, the addition of new nucleotides to a generated molecule of DNA happens with the 3'-end of the nucleotide chain (figure 2.3a). Chain terminators are dideoxy nucleotides, ddNTPs, that cannot be further extended. These nucleotides aren't able to add a new nucleotide on the 3'-end, as they have not the needed oxydrilic group (figure 2.3b).

2.1.3 Sanger method: the first one

The first method ever used to sequence DNA was designed by Frederick Sanger. The Sanger manual sequencing system consists in an *in vitro* process, which is described in figure 2.4, also named as

2.1. HISTORY OF SEQUENCING

Figure 2.4: The Sanger's method process



"primer extension" method. It is performed over a single-filament DNA sample, and it uses the chain terminators nucleotides, a type for each nucleobase: ddATP, ddGTP, ddTTP, ddCTP.

The reaction is done inside four different reactions tubes, each containing the sample DNA to be reproduced, a DNA polymerase, the normal nucleotides and one of the four possible chain terminators. The chain terminators are marked with sulfur-35, a radioactive atom. In each tube, the corresponding dideoxy-nucleotide was used with a concentration 10 times lower than the other "normal" nucleotides.

From the polymerization reactions, several molecules of DNA were produced, with different length: each replicative cycle is in fact terminated after the addition of a chain terminator nucleotide.

To reconstruct the initial DNA sequence, a long PAGE gel was prepared, with high concentration of urea ($6 - 7M$) to avoid the coiling of the DNA single-filaments. To run the gel, high voltages were required and it had to be highly resolute, as DNA's fragments are different only for a nucleotide. It was needed to do an auto-radiography of the gel in order to see the bands, in order to differentiate the fluorescent signals.

To read the sequence you have to start from the shortest fragments, at the end of the gel, and carefully go up along the gel, looking for the first presence of a band in one of the four runs.

Past procedures: In the past, the Klenow's fragment (part of the 1st DNA polymerase) was used to perform the Sanger method, and the DNA to be sequenced was inserted inside the genome of an M13 phage.

2.2. DEVELOPMENT OF SEQUENCING MACHINES

2.1.3.1 Automatic sequencing

It does not use the radioactive signals, instead it uses fluorescent proteins. Several versions were developed after modifications of the Sanger method, in this order:

1. fluorescent primers marked with a single fluorochrome.
2. four aliquotes of the same primer were used marked with four different fluorochromes, able to emit different fluorescences.
3. four different fluorochromes were used to mark the single ddNTPs

Thanks to the use of 4 different fluorochromes, it was possible to use a single electrophoretic lane to carry the sequencing reaction. For this type of sequencing also, a cyclic replicative reaction was performed, with this procedure, made possible by using a thermal cycler:

1. **Denaturation at 95°C** of the DNA to be sequences
2. **Annealing at 50 – 70°C** of the primer specific to one of the two filaments
3. **Extension at 72°C** by using a *Taq*-polymerase. The use of the *Taq*-polymerase makes it possible to avoid the formation of coiled structures in the DNA molecule to be sequenced.

Traditionally, also in this type of sequencing it is used a long PAGE gel, as the Sanger's one, but with a great difference: all the ddNTPs are inserted in the same electrophoretic run, and after it, it is not needed an auto-radiography. Fluorescence, instead, can be triggered simply by irradiating the DNA molecules synthesized, which produce different fluorescences with different wavelengths.

More usefully, this sequencing method is performed by using a capillary, filled with a synthetic polymer, with the same function of polyacrylamide.

At the end of the analysis, it is produced an electropherogram, with a color depicting the probability of each nitrogen base in each position. The production of the electropherogram is made better thanks to algorithms to boost signal/noise ratio, to correct the dye-effects, and other effects that generate systemic errors.

Regarding the Sanger machines in use, the upgrades viewable in figure ???. Based on the same technology, new machines were developed and it was obtained a 1000 fold improvement.

When talking about the **human project**, it is important to specify that most of the job was done by using the Sanger automatic sequencing method.

2.2 Development of Sequencing Machines

The way you can get to the point could be different based on technology and the wanted output, comprised quality.

SOLID gave only 35-75 sequenced bases, and it is not used anymore; Sanger sequencing, the capillary, can give up to 1000 read length, with a low output; on MINION it was possible to sequence an entire genome of *E. coli*. A plethora of sequence machines are available today. None of the machines are able to sequence DNA from a sample of blood, some things have to be done. Nowadays machines producing bigger outputs of short reads are preferred.

2.2. DEVELOPMENT OF SEQUENCING MACHINES

Figure 2.5

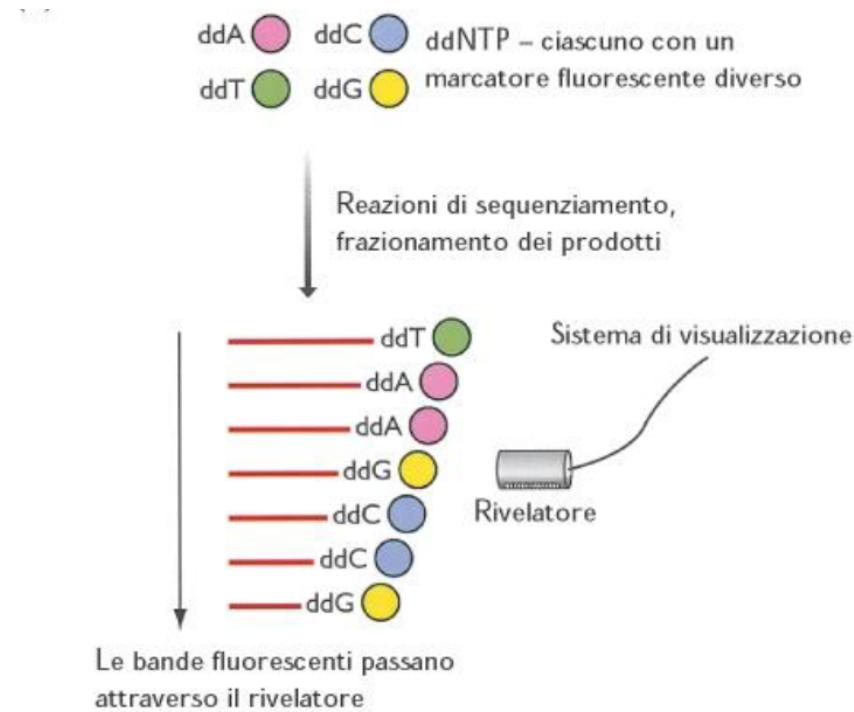
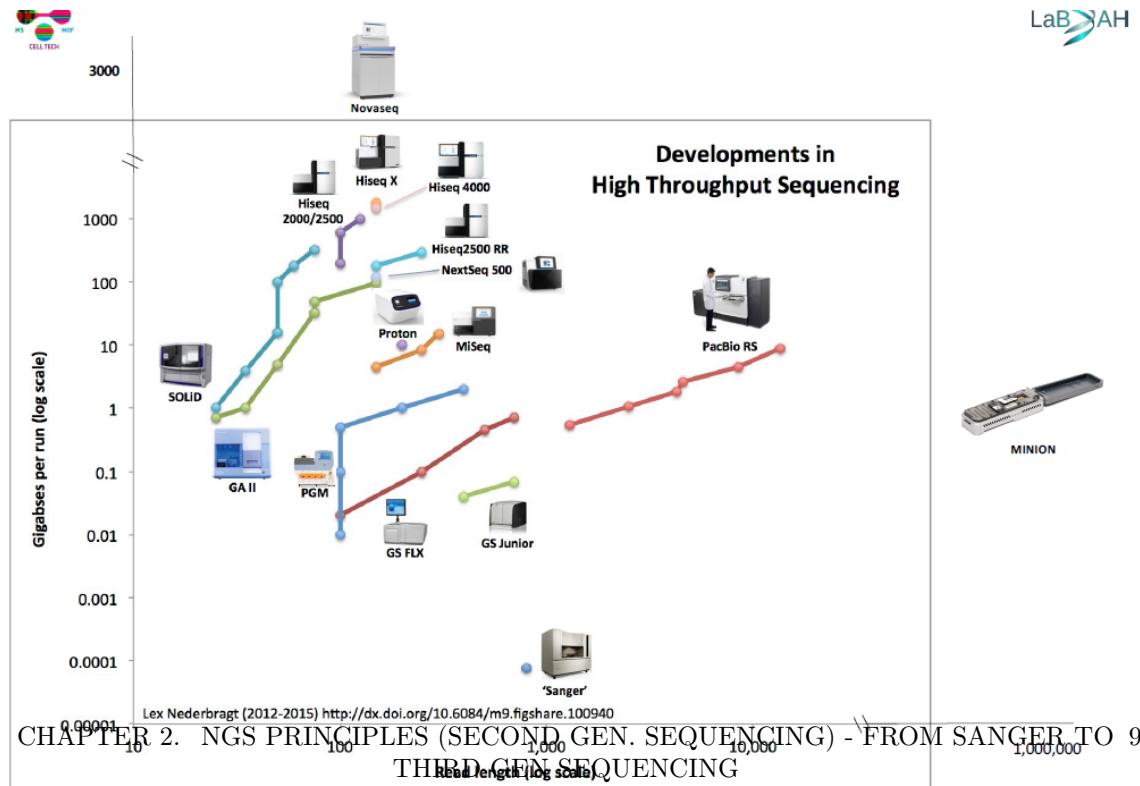
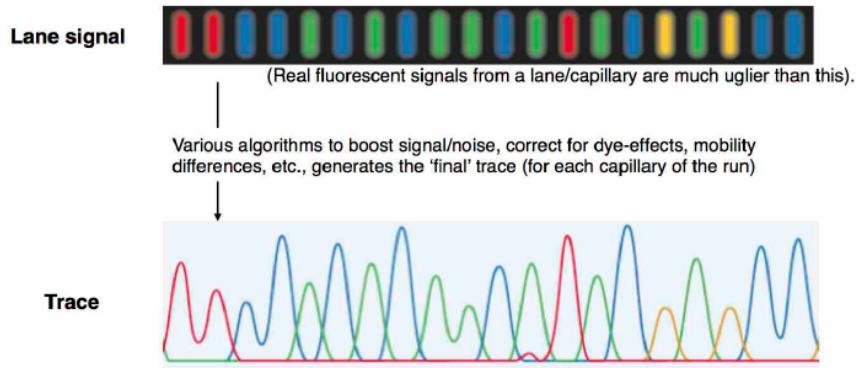


Figure 2.8: It can be noticed how recent developments had the scope of increasing the output data



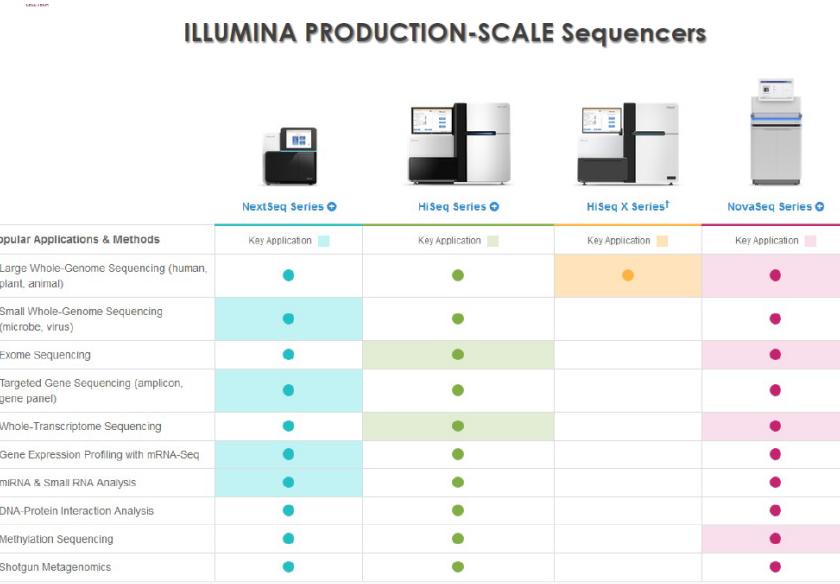
2.3. NEXT GENERATION SEQUENCING NGS

Figure 2.6



At the top of the market nowadays there are ILLUMINA machines, that use sequencing by synthesis method, NovaSeq is the biggest one. They need to amplify the signal through clusters formation.

Figure 2.9



2.3 Next Generation Sequencing NGS

The NGS protocol requires 3 steps, that are:

1. **Sample preparation:** series of fragments added

2.3. NEXT GENERATION SEQUENCING NGS

Figure 2.7: The implementation of capillary sequencing machines gave the possibility to make more runs than with the others. A ~1000 fold productivity increase was allowed

		
Radioactive polyacrylamide slab gel Low throughput, labor intensive	AB slab gel sequencers (370, 373, 377) Fluorescent sequencing 1990-1999 6 runs/day 96 reads/run 500 bp/read 288,000 bp/day	AB capillary sequencers (3700, 3730) 1998-now 24 runs/day 96 reads/run 550 – 1,000 bp/read 1-2 million bp/day

2. **Clonal amplification:** which is needed to replicate fragments attached to the solid surfaces, since machines are not sensible to single molecules
3. **Sequencing:** ILLUMINA sequencing is one of the techniques used to obtain sequence data nowadays

Tools pertaining to the 3rd generation are those that permit to read a molecule without replicating it.

2.3.1 Fragments/Library preparation

Most of the sequences sequenced are fragmented in short read sequences, since most of the machines today used aren't able to sequence reads longer than some hundreds of nucleotides.

The fragments obtained have to be prepared for the sequencing process, through a process called **tagmentation**. The obtained fragments are shown in the figure. Those fragments are provided with one or two indexes, called also barcodes, two sequencing primer binding sites and regions complementary to the oligonucleotides present in the chamber (see in Clonal amplification chapter). The fragments' length has to be checked, depending on the scope of the process. Indexes are needed to run sequencing process on multiple samples, they are needed to distinguish those; when they are two, they permit to distinguish also the 2 types of sequencing that are performed, forward and reverse.

P5 and P7 are the oligos needed to attach fragments to ILLUMINA sequencing machines.

2.3.2 Clonal amplification and ILLUMINA sequencing procedure

Clonal amplification are necessary to amplify the signal from each single fragment. ILLUMINA machines make use of clusters to sequence DNA. Clusters are a group of DNA strands positioned closely together and generated from a single DNA filament. Generally, Each cluster represents thousands of copies of the same DNA strand in a 1–2 micron spot (figure 2.11).

2.3. NEXT GENERATION SEQUENCING NGS

Figure 2.10: Figure representing the a good prepared fragment, it has two indexes, two sequencing primer binding sites and regions complementary to the oligonucleotides present in the chamber

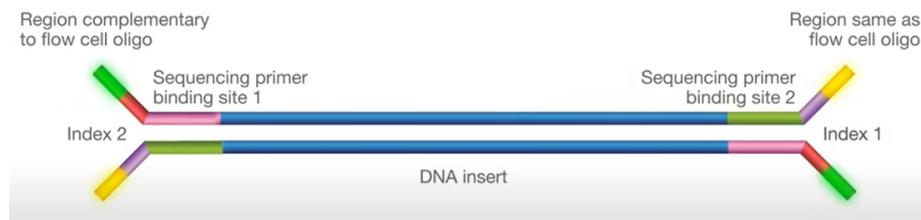
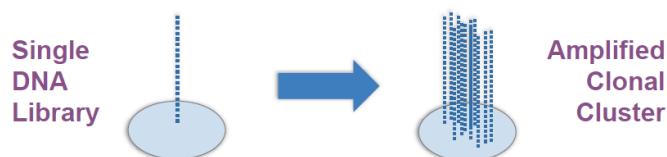


Figure 2.11



On patterned flow cells, the clusters' formation location could be known or not, in the first case it is said that there is a "Rigid registration" (figure 2.12).

ILLUMINA sequencers normally use slides of glass, named flow cells, and the fragments to be sequenced are able to flow over the channels. The temperature inside the cells can be changed to produce ligations or separations. The surface of the cells is functionalized with a series of oligos complementary to library adapters.

Two kinds of flow cells, the *patterned* flow cell permits to create clusters in specific positions, inside nanowalls, contrarily to *Random Flow Cell* which instead have randomly positioned clusters.

Once the fragments are made flow over the chambers, they can bind only to p5 or p7 (ILLUMINA oligos), the two oligos functionalizing the plate. Once the fragments are attached to the surface, using temperature and solvents flows you can control the sequencing process. To see the entire procedure: (procedure on Youtube: video about ILLUMINA sequencing). In the video, it is shown the two index sequencing process.

2.3. NEXT GENERATION SEQUENCING NGS

Figure 2.12

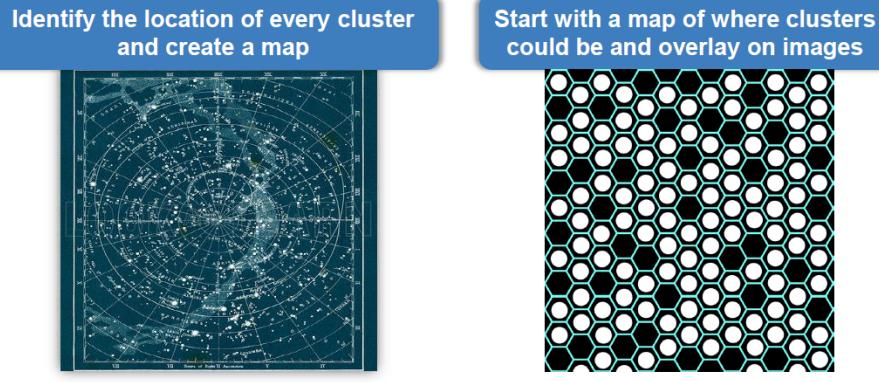


Figure 2.13

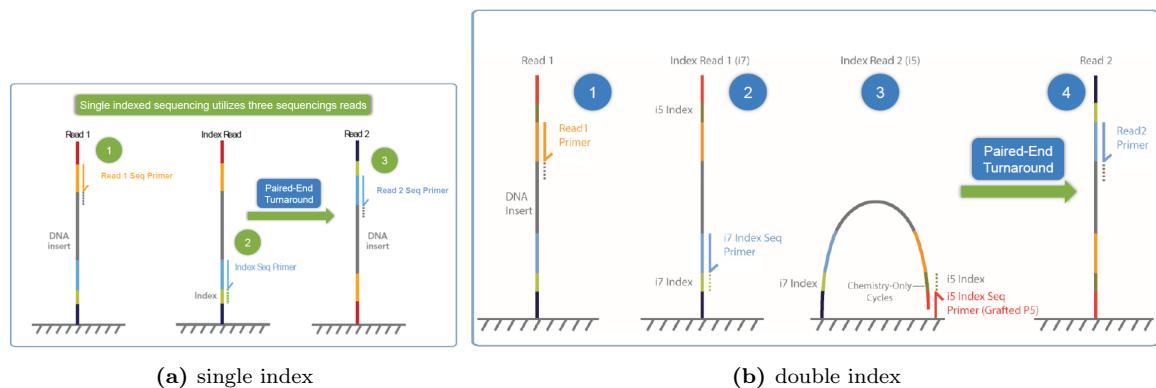
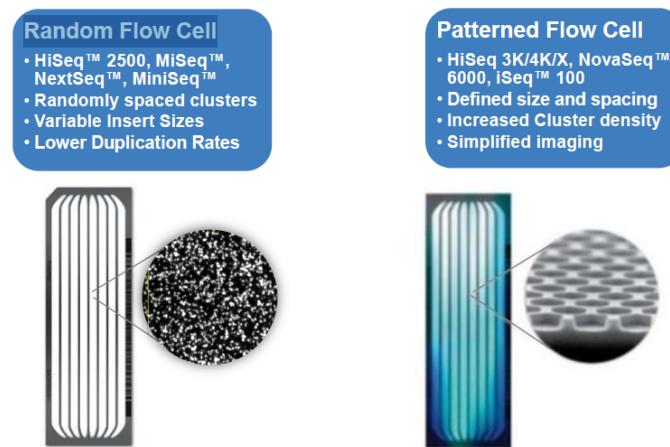
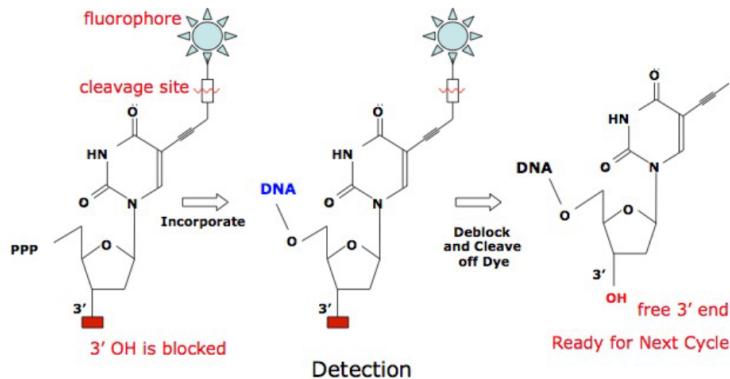


Figure 2.14: Single/double index for ILLUMINA sequencing

2.3. NEXT GENERATION SEQUENCING NGS

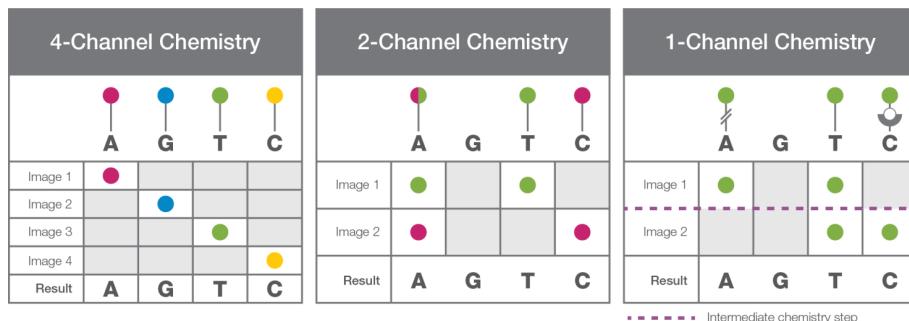
With 1 and 2 indexes the sequencing process actually differs, as shown in figure 2.14

Figure 2.15



To perform the sequencing process, ILLUMINA machines utilize *reversible terminators* (figure 2.15). They permit a real time analysis of the sequencing by synthesis reaction, and because of this they are different from the Sanger method. Fluorophores are reversible, they can be cleaved to eliminate the light signal.

Figure 2.16



To perform their activity, ILLUMINA sequencers could be of 3 different types: 4-Channel, 2-Channel or 1-Channel, depending on the number of fluorescent molecules used. In the case of the 4-Channel technology, 4 images are taken in each cycle, and each cluster appears in only one of four images 2.16. 2-Channel technique is used by some sequencing machines, like NextSeq 550, MiniSeq, NovaSeq 6000

4-Colors base calling is needed to make the true signal the purest, and after, The base with the highest intensity becomes the called base for that cluster. In the case no base is clearly related to a position, *N* is the result.

The reading process could be done in two ways: through single reads, on a single extreme of the fragments, or paired-end, on both the extremes. The second one in particular gives structural and 2 sequence informations.

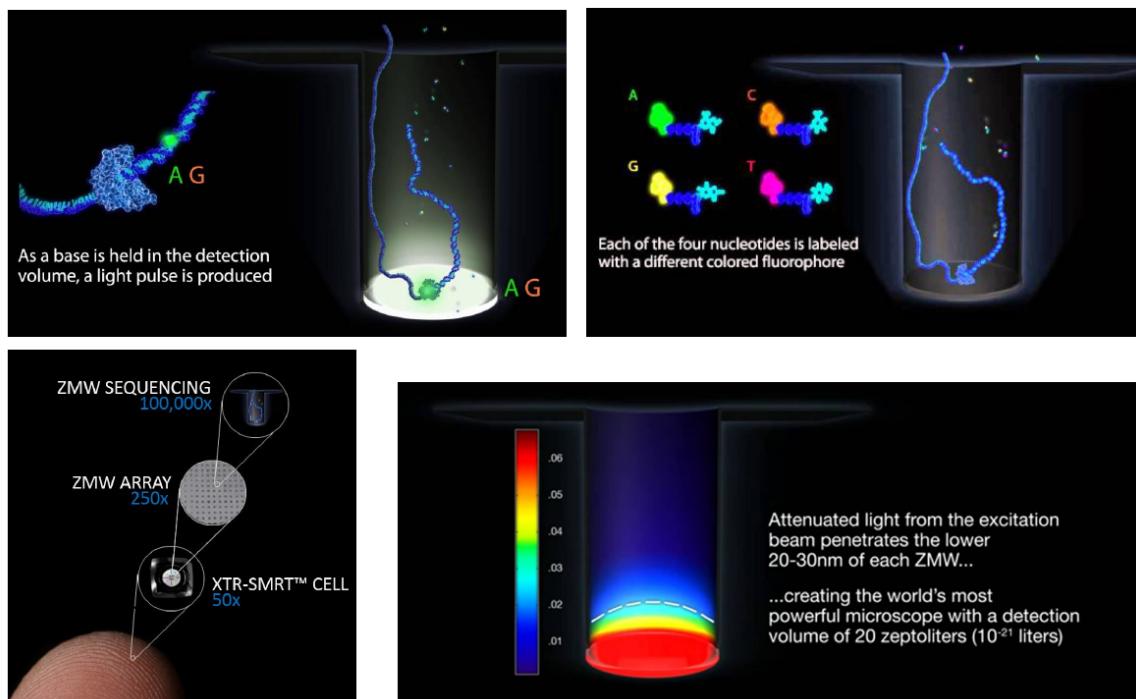
2.3. NEXT GENERATION SEQUENCING NGS

2.3.3 Pacific Bioscience (PacBio)

the long DNA filament to be sequenced is attached to a polymerase, over the surface of a SMRT (Single Molecule Real Time) cell. This cell is really small, and at each nucleation process a light signal is emitted. The produced light is not able to get out of the walls, and its duration is extremely restricted. Sequencing, also in this case, is made by sequencing, and the main advantage consists in the possibility of sequencing really big DNA molecules.

In the video PacBio procedure it is briefly shown how the process works and what are the strategies used to reduce errors.

Figure 2.17



2.3.4 Nanopore sequencing

Intramembrane proteins are used, sequence detected through the passage of DNA nucleotides, which produce different voltage changes. During the first periods, this type of instruments gave great amount of errors, nowadays the technique is improving. Nanopore Sequencing input fragments. the order of the contigs cannot be established unless we have a connection. De novo assembly without prior information about the sequence of the organism. Sometimes it could be useful to use available reference sequences. Artifacts have to be discriminated. It is possible to generate a consensus sequence.

Sequence mapping aligning: find dna sequences dsimilarities between query sequences and reference sequences.

2.4 Sequence mapping

Query short sequence and reference is much longer. Could be deletions, insertions, substitutions. THe first two can be longer than 1. You would search for the best match or the best matches, or all matches, the latter is very computationally demanding.

2.4.1 Coverage

Average number of reads representing a given nucleotide in the reconstructed sequence. Some positions could have 0 coverage.

$$Cov = \frac{NxL}{G}$$

where G: length of the genome, N: number of reads, L: average read length.

The coverage is an essential parameter, more coverage means more reads to do.

Exercise: I want to sequence a genome which I know is 3M bases. I can use Illumina single.end sequencing with read length of 100 nt.

From sequence comparison it is possible to obtain new informations.

Simplest aligning algorithm: start from 0 position and slide along the sequence. It doesn't consider indels.

Global alignments focus on total sequence, while local alignment only in fractions of the total sequence.

Evolutionary distance doesn't reflect precisely evolutionary distances when the observed percent identities are low.

2.4.2 Smit-Waterman algorithm

"Identification of Common Molecular Subsequences", how many changes to make two sequences identical. they wanted a solution such that gives the best matches.

We have to find best substrings. It has to be built a big matrix, consider a score, used when nucleotides match or not. Populate the matrix, in each cell a number described by the formula. Select values of match and mismatch and a penalty value for gaps.

more than one best solutions.

This is not used anymore, as it would result in too much computational expense.

2.4.3 Needleman-Wunsch algorithm

Optimal align algorithm. initialized with 0, 0 is not possible anymore for iteration. Termination in the bottom right corner.

If we want an accurate alignment, these algorithms could in some cases used. People developed different strategies algorithms, FastA, Blast, BLAT, BWA, BowTie2. THe last two are optimized for short reads. heuristic means that you are considering a threshold, if a certain level of identity is not respected, than the program doesn't retrieve any output.

2.5 Blast

very used nowadays, still working. Allowed sequence matching to everyone through affordable computers. It can be run online, they set very high seeds, to avoid computing too much

2.6. SPACED SEED ALIGNMENT

Find k-mer matches. Once you find some good seeds, it is possible to extend the matches to see if there are some more extended regions. The evaluations are made with algorithms inspired on...

k is generally between 5 and much larger numbers.

several BLAST programmes are available depending on the type of need.

E-value: the number of distinct alignments with a score equivalent to or higher than S, that are expected to occur in a database search by chance

$$E = Kmne^{-\lambda} \quad (2.1)$$

There is a relation between P-value and the E-value. P-value: the probability to obtain by chance another score with an equal or better score.

$$\frac{E}{mn} \quad (2.2)$$

where

1. Parameter K and λ depend on the substitution matrix and on the gap...

2.5.1 Scoring matrices

A protein which changes an AA with another not always changes it with a completely different AA. Points are established through biochemical evaluations.

2.6 Spaced seed alignment

They are not used anymore. Reads are cut down in small seeds. The seeds are stored in an index, and then seeds are searched inside the sequence.

It is possible to have SNPs. The look up table is actually enormous in human genome
Maq, SOAP, MOSAIK, Novoalign are based on this approach.

2.7 Burrows-Wheeler transform

Reversible permutation of the characters in a text. Compress a file, it takes a lot of time to compress a file.

Fast reconstruction using the LF property: the ith occurrence of character X...
Several ways to decompress com

Chapter 3

Sequencing Data

hard drive with the sequencing data

3.0.0.0.1 a biological problem, which sequencing machine is optimal?

- (no problem with) repeats
- read line
- throughput ($\frac{\text{Millions of reads}}{\text{run}}$)
- library preparation compatibility
- coverage
- cost: of the sequencing of the machine itself. $\frac{\text{Reagent cost}}{\text{run}}$, $\frac{\text{cost}}{\text{Mb}}$, $\frac{\text{cost}}{\text{run}}$, service contract
- accuracy: multiple types of errors: Indel, Substitution, CG deletions (deletions o full CG), A-T bias. The error rate takes in account all the primary errors of the specific instrument.
- speed / run-time
- contamination

GA II was the first version of illumina. SOLID was a sequencing machine which is not anymore used. Roche454 responsible of revolution in 2007, not used anymore. MiSeq is also used today. Long read sequencing MinION and pacBio are probably next generation. pacBio has very long reads and close to single molecule sequencing. Ion-Torrent is not anymore used. MinIon and PacBio gives real time sequencing. 454 quite high number of reads but the throughput low.

Going towards high throughput.

Illumina NovaSeq is the biggest in the market. very high throughput need.

ILLUMINA has very small machines, not expensive

MinION is portable, wet-lab free, real time sequencing. It was used in Ebola pandemic- Used for sequencing DNA, RNA, proteins. The read length follows a Gaussian curve, as there can be also really long reads.

After base calling, which is the process through machine converts signals in digits representing the bases. It is made a FASTQ format file. other machines generate other files, but all can at least

convert their output in FASTQ. FASTQ contains the sequencing reads and the quality of each base. FASTQ needs to be stored compressed.

Phred- the base calling program, are based on algorithms, to check statistically a sequence. The phasing noise ϕ , the intensity skew the intensity of the signal of the closest nucleotide.

signal decay δ

on ILLUMINA there can be mixed clusters.

Boundary effects: image if zoomed in it is not really possible to give perimeters to the images of the signals. Overclustering makes it very difficult to separate different nucleotide produced signals. Underclustering and Overclustering, optimal clustering in between.

Base-caller has to be performing. By sequencing a known sequence it is possible to evaluate the base caller.

3.0.1 FASTQ format

1. "@" followed by a sequence identifier
2. the sequence
3. "+", optionally followed by a sequence identifier
4. The quality scores: several symbols.

Base quality:

$$Q = -10 \log_{10} p$$

p = probability of error

Phred+33 transforms integers with characters. Each number can be told in an integer.

FASTA format

1. ">" followed by a sequence identifier
2. The sequence

Transform FASTQ in FASTA format file

3.0.2 Quality control

The image under has not a good distribution of read lengths in bp. The first positioned reads are normally the best in terms of quality, Going over the quality lowers. The optimal case would be when quality remains constant. Bugs in the base caller could also variate the quality plot. Green, orange and red regions.

Adapters have to be considered to maintain good quality sequences. Removing them means also to reduce the dimension of the output.

$$x = 1\text{ciao}$$

Nanopore data maintain the quality of the sequencing. PacBio seems to behave in really similar way.

Distribution of k-mers: k-mers are a lot in the small numbers region, because of the sequencing errors.

Low-complexity artifacts: low complexity and entropy and high compression.

Chapter 4

Staphylococcus aureus