



LEIDEN UNIVERSITY MEDICAL CENTER

General Introduction

Michiel van Galen and Jeroen F. J. Laros
Department of Human Genetics
Center for Human and Clinical Genetics



A selection of the landmarks in sequencing.

Year	Landmark
1953	Discovery of the structure of the DNA double helix.
1977	The first complete DNA genome sequenced (Φ X174).
1986	First semi-automated DNA sequencing machine.
1995	First complete genome of a free-living organism.
2001	A draft sequence of the human genome is published.
2004	454 Life Sciences markets a parallelized version of pyrosequencing.

High throughput parallel sequencing.

- Developed started in the 1990s.
- Became available on the market in 2004.

Nowadays there are three major platforms available

- Solid sequencing.
- 454 pyrosequencing.
- Illumina (Solexa) sequencing.

SOLiD: Sequencing by Oligonucleotide Ligation and Detection.

- Manufacturer: Life Technologies.
- Commercially available since 2008.
- Per cycle, 2 bases are read.
- Each base is read twice.

Pros:

- Single base read errors can be detected.

Cons:

- Software has to be adapted to work in *colour space*.
- The first nucleotide *must* be known.

454 pyrosequencing

Long reads with the Roche 454.

- Manufacturer: 454 Life Sciences (Roche).
- Commercially available since 2005.
- Per cycle, a variable number of bases are read (monomer stretches).
- Used in forensic science.

Pros:

- Long reads (400-500 base pairs, 1000 announced).

Cons:

- Deals poorly with monomer stretches.

The Illumina Genome Analyser II (GAII).

- Manufacturer: Illumina, Inc.
- Commercially available since 2005.
- Per cycle, one base is read.
- Reads up to 100×2 base pairs.
- Takes about 8 days.
- Produces about 40 Giga bases per run.

Pros:

- Does paired end sequencing.
- Cheap.

The Illumina HiSeq.

- Manufacturer: Illumina, Inc.
- Commercially available since 2010.
- Per cycle, one base is read.
- Reads up to 150×2 base pairs.
- Takes about 8 days.
- Produces about 150 Giga bases per run.

Pros:

- Even higher throughput.

We have the following sequencers at our disposal.

First generation, mainly used for short sequences and validation.

- 3 Sanger sequencers: (2× 3100 and 3730).

Next generation, high throughput.

- 2 Illumina GAI.
- 1 Illumina HiSeq (got it last week).
- 1 Roche 454.

Next (and a half) generation.

- 1 Helicos.

Paired end sequencing

By sequencing two ends of a molecule, mapping can be improved.

- If one end maps more than once, the second end can be used to uniquely determine the position.

After the first read is done:

- Sequencing stops and chemicals are refreshed.
- A second *sequencing primer* is added.
- The sequencer is started again.

Note that:

- It is very important that the flow cell is not moved.
- Typically, the second read is shorter than the first one.

Barcoding is a way to tag samples, in order for them to be pooled in one lane.

Afterwards, we use the barcode to split the data.

- Usually added in the PCR step.
- Designed in such a way, that single errors can be corrected.
 - The *edit distance* between two barcodes is at least three.
 - Introducing one error (insertion, deletion or substitution) will result in a edit distance of one.
 - Since the distance to other barcodes is still at least two, we can assign the barcode anyway.
- Can be used as an additional quality control.

True single molecule sequencing.

Pros:

- No amplification.
- Detection of mixed samples (forensics).

Cons:

- Short reads (32).
- Suffers from *dark nucleotides*.
- Has a high error rate.
- No barcoding.
- No paired end (yet).

Characteristics:

- Single molecule.
- Long reads (several kilobases).

Applications:

- De novo assembly.

Companies:

- Pacific Bio.
- Ion Torrent.

We start with a basic introduction to Linux.

- Why do we use / need Linux?
- Using the *command line* and NGS tools (building blocks).
- Connecting to other machines.

Furthermore, we focus on pipelines.

- Combine the building blocks into a pipeline.
- Common pipelines.
 - Variant detection.
 - Expression analysis.
- Using *Galaxy*.

Michiel van Galen
Jeroen Laros

<http://www.mutalyzer.nl/NGSIntroCourse/>