Introduction to NGS technologies

Joaquín Panadero Romero

2nd March 2016
1. Basics on the NGS technologies
2. Comparisons across NGS platforms
3. Computing infrastructure for NGS analyses
4. Tools for data analysis
Basic on NGS technologies

Millions of DNA molecules sequenced simultaneously

- Personalized medicine
- Genetic diseases
- Clinical diagnostics

**Types:**

- Sanger
- Pyrosequencing
- Sequencing by synthesis
- Sequencing by ligation
- Ion-Semiconductor sequencing
Used nowadays in:
- Routine sequencing applications
- NGS data validation

Multiple DNA fragments covering each base position
DNA fragments move according to their size.

Light detected shows the base added at each position.
5. Common among NGS technologies

1. Sample preparation
   - cDNA fragments ligated to adaptors at both ends
   - Amplification based on PCR bridges or bead emulsion

2. Sequencing machine
   - Method 1: Bridge PCR
   - Method 2: Emulsion PCR

3. Data output
   - Sequencing output is provided in clusters
   - Raw data presented on DNA chips
Common among NGS technologies

1. Prepare genomic DNA sample
2. Attach DNA to surface
3. Bridge amplification

4. Fragments become double-stranded
5. Denature the double-stranded molecules
6. Complete amplification

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate. Denaturation leaves single-stranded templates anchored to the substrate. Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
Common among NGS technologies
Overview

- Large reads lengths generation
- High reagent cost
- High error rate over strings of 6+ homopolymers
Sequencing by synthesis

Overview
- Overcomes homopolymer issue due to terminated nucleotides
- Increased error rate with increased read lengths
Sequencing by ligation

- 16 8-mer oligonucleotide probes
- Ligase connects the adaptor to the sequence
- First base: A, C, G, T
- Second base: A, C, G, T
Sequencing by ligation

5 x 7 ligation cycles. Each primer hybridizes one base back

**Overview**

- Oligonucleotide probes used rather than DNA Polymerase
- Very short read lengths
Ion - Semiconductor sequencing

Beads are attached to semiconductor transistors

Each time a nucleotide is added, one H+ is released

Semiconductor transistor detects changes on pH solution
Overview

- Similar to pyrosequencing, but measures the release of H+ instead of pyrophosphate
- Most cost-effective and time-efficient
Examples of NGS systems

- Illumina MiSeq
- Illumina HiSeq
- Solid system
- Ion Proton
- 454 Sequencer
## NGS comparison

<table>
<thead>
<tr>
<th>Coverage of genome per run</th>
<th>Human</th>
<th>Mouse</th>
<th>Plant</th>
<th>Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrosequencing</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>151</td>
</tr>
<tr>
<td>Sequencing by synthesis</td>
<td>455</td>
<td>536</td>
<td>11k</td>
<td>323k</td>
</tr>
<tr>
<td>Sequencing by ligation</td>
<td>97</td>
<td>114</td>
<td>2k</td>
<td>69k</td>
</tr>
<tr>
<td>Ion semiconductor sequencing</td>
<td>3</td>
<td>4</td>
<td>74</td>
<td>2k</td>
</tr>
</tbody>
</table>
Applications

- Whole genome sequencing
- Variant Calling
- RNA-seq
- De novo sequencing and assembly
- Chip-seq
- Methyl-seq
- Metagenomics
Sequencing costs

Full Genome Sequencing & The Genetic Revolution
Cost per Human Genome vs Total Number of Genomes Sequenced

- Cost per Human Genome for Full Genome Sequencing
- Total Number of Human Genomes Sequenced

Industry data from public online sources
Dashed lines represent extrapolations based upon current trends
- In NGS we have to process really big amounts of data, which is not trivial in computing terms
- Big NGS projects require supercomputing infrastructures

thus

we can tackle such amount of data by using specific hardware combined with software capable to deal with data generated
Requirements:
- Conditioned data center (server rooms)
- Computing cluster (racks)
- Many computing nodes (servers)
- High performance and high capacity storage
- Fast networks (10Gb ethernet, infiniband...)
- Skilled people in computing (sysadmins and developers)
Computing cluster and storage

Distributed memory cluster
- 8 or 12 cores per node
- At least 48GB RAM per node

Fast networks
- 10 Gbit, infiniband...

Batch queue system
- sge, slurm, condor, pbs
What do we want to store?

- Raw data (fastq)
- Processed data (fastq, bam, sam, vcf)
- Final results (txt, excel...)

How many storage resources?
For how long?
Sequencing instruments
- 10 Illumina HiSeq2000

Informatics infrastructure
- 850 core cluster
- 7.5 petabytes, lustre filesystem
- 10 x 10 Gb link with MareNostrum
Sequencing instruments
- Illumina HiSeq
- AB Solid System
- Ion Torrent

Informatics infrastructure
- 20576 cores cluster
- 17 PB (petabytes)
Alternatives: cloud computing

**Pros**
- flexibility
- you pay what you use
- don't need to maintain a data center

**Cons**
- transfer datasets through the internet is slow
- lower performance
- privacy and security concerns
- more expensive for big and long term projects
<table>
<thead>
<tr>
<th>Tools on NGS data analysis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quality</strong></td>
<td><strong>FastQC</strong></td>
</tr>
<tr>
<td><strong>Trimming</strong></td>
<td><strong>cutadapt</strong></td>
</tr>
<tr>
<td><strong>Assembly</strong></td>
<td><strong>abyss, velvet, ...</strong></td>
</tr>
<tr>
<td><strong>ORFs prediction</strong></td>
<td><strong>glimmer, augustus, ...</strong></td>
</tr>
<tr>
<td><strong>Annotation</strong></td>
<td><strong>Blast2GO</strong></td>
</tr>
<tr>
<td><strong>Mapping</strong></td>
<td><strong>BWA, bowtie, hpgaligner</strong></td>
</tr>
</tbody>
</table>
## Tools on NGS data analysis

<table>
<thead>
<tr>
<th>Category</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential expression</td>
<td>babelomics, cuffdiff, bioconductor</td>
</tr>
<tr>
<td>Variant calling and Variant annotations</td>
<td>GATK, samtools, Annovar, BiERapp</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>qimme, mothur</td>
</tr>
<tr>
<td>Methylation</td>
<td>bismark</td>
</tr>
<tr>
<td>Functional profiling</td>
<td>babelomics</td>
</tr>
<tr>
<td>Path signaling</td>
<td>hiPathia and Pathact</td>
</tr>
</tbody>
</table>
NGS Data Analysis Pipeline

- Sequence preprocessing
  - Alignment
    - Visualization
      - Variant calling
        - Variant annotation
          - Prioritization
      - RNAseq processing
        - RNAseq data analysis
          - Functional analysis
NGS Data Analysis Pipeline

1. Sequence preprocessing
2. Assembly
3. Contig filtering
4. Annotation
5. Alignment
6. Data analysis
NGS Data Analysis Pipeline

1. Sequence preprocessing
2. Alignment
3. Clado counting
4. Results
THANKS