GALAXY INITIATION

A. Lermine

U900 Institut Curie, INSERM, Mines ParisTech
How does Next-Gen sequencing work?

1. DNA fragmentation
2. Size selection and clonal amplification
3. Massive parallel sequencing

ACCGTTTGCCG…

$10^6$-$10^9$ short reads功课ioned in FASTQ format
Hi-seq, SOLiD, PGM, ..., What does it mean?

<table>
<thead>
<tr>
<th>Platform</th>
<th>Provider</th>
<th>Reads Number (M)</th>
<th>Max Reads Size (bp)</th>
<th>Throughput (Gb)</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS Flex</td>
<td>Roche</td>
<td>1</td>
<td>700</td>
<td>0.7</td>
<td>8d</td>
<td>Base</td>
</tr>
<tr>
<td>Hiseq 2000/2500 Normal mode</td>
<td>Illumina</td>
<td>3000</td>
<td>2x100</td>
<td>600</td>
<td>11d</td>
<td>Base</td>
</tr>
<tr>
<td>Hiseq 2500 rapid mode</td>
<td>Illumina</td>
<td>600</td>
<td>2x150</td>
<td>120</td>
<td>40h</td>
<td>Base</td>
</tr>
<tr>
<td>MiSeq</td>
<td>Illumina</td>
<td>15</td>
<td>2x250</td>
<td>8.5</td>
<td>40h</td>
<td>Base</td>
</tr>
<tr>
<td>SOLiD</td>
<td>LifeTech</td>
<td>1400</td>
<td>75-35</td>
<td>150</td>
<td>25d</td>
<td>Color</td>
</tr>
<tr>
<td>PGM 314</td>
<td>Ion Torrent</td>
<td>0.5</td>
<td>400</td>
<td>&gt;0.01</td>
<td>2-4h</td>
<td>Base</td>
</tr>
<tr>
<td>PGM 316</td>
<td>Ion Torrent</td>
<td>2</td>
<td>400</td>
<td>&gt;0.1</td>
<td>2-4h</td>
<td>Base</td>
</tr>
<tr>
<td>PGM 318</td>
<td>Ion Torrent</td>
<td>4</td>
<td>400</td>
<td>&gt;1</td>
<td>2-4h</td>
<td>Base</td>
</tr>
<tr>
<td>Proton</td>
<td>Ion Torrent</td>
<td>60-80 M</td>
<td>200</td>
<td>&gt;12</td>
<td>2-4h</td>
<td>Base</td>
</tr>
</tbody>
</table>
Let’s start after sequencing ...

- A **raw** data file containing **millions** of reads (A/C/G/T/N sequence + base qualities):
  - **IonTorrent PGM**: FASTQ, SFF, Unmapped BAM of reads of different sizes
    → SFF and Unmapped BAM contain specific Ion Torrent flowspace information (helps improve the accuracy during the mapping and variant calling steps)
  - **Illumina**: FASTQ of same size reads

- Reads can be **single-end** (one end of the fragment sequenced) or **paired-end** (both ends sequenced)
Galaxy Workflow

- **Reads (Fastq)**
  - Conversion to Galaxy Format (Groomer)
  - Quality Control (FastQC)

- Reference Genome (Fasta)

- Mapping (Bowtie2)
  - PCR duplicates Marking (MarkDup)
    - Not for PGM data

- Mapping statistics With flagstat
Galaxy Workflow

Conversion to Galaxy Format

Groomer

Reference Genome (Fasta)

Mapping

Bowtie2

Mapping statistics With flagstat

PCG duplicates Marking

Not for PGM data

MarkDup

Quality Control

FastQC

Reads (Fastq)
Two available datasets on Galaxy

1. Open your web browser and go to « http://galaxy.sb-roscoff.fr »

2. In the top menu, click on « Shared Data » then « Data libraries »

3. Click on « TP-INITIATION »:
   - Chr4.fastq: raw reads file
   - Chr4.fasta: reference sequence for alignment
Sequence Quality Encoding (FASTQ)

- Extension from traditional FASTA format

- Each block has 4 elements (in 4 lines):
  
  - Sequence name (read name, group, etc…)
  
  - Sequence
  
  - + (optional: sequence name again)
  
  - Associated quality scores (phred-scaled) : different encoding possible

- Example record:

  - @FCD19MJACXX:2:1101:1735:1993#GTTCGACA/1
  
  - NGAGGCTGAGGCGGGCAGAGGTCAGGAGATCGAGACCATC

  - +

  - BP\cccccc]ceecheheeZbe_cZbd_dbbdd\aXab_`b
Sequence Quality Encoding (FASTQ)

- The base calling (A, T, G or C) is performed based on Phred Scores.

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of Incorrect Base Call</th>
<th>Base Call Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

- Phred scores provide $\log_{10}$-transformed error probability values
  - If $p$ is probability that the base call is wrong the Phred score is

\[
Q = -10 \log_{10} (P) \iff P = 10^{-Q/10}
\]
FASTQ format conversion

1. Rename your history to « TP Initiation » by clicking on «Unnamed history »

2. In the left panel, click on « FASTQ Groomer » under the NGS: QC and manipulation section to convert your FASTQ into FASTQ Sanger Format

3. Click on « Execute » to launch the conversion
FASTQC: Quality Control on raw data

- **Sequence Length Distribution**: sequencers produce same (Illumina) or different (PGM) read length. This metric helps identify abnormal read length.

- **Sequence content**: % A/C/G/T ; %GC; %N at each position in the read
  
  → Proportion biased for targeted sequencing

- **Quality score**:
  
  1. **Per base**: identify base calls with low quality (commonly towards the end of a read)
  
  2. **Per sequence**: to see if a subset of your sequences have universally low quality values

- **K-mers content**: a k-mer is a motif of length k observed more than once in a sequence (repeats : ACACAC ; spaced occurrences : tccGAGGaaggGAGGaag)

- **Over-represented sequences**: highly duplicated sequence in your library (primer, adapter..)
FASTQC : FASTQ Quality Control

1. In the left panel, click on « FASTQC: Read QC »

2. Select the FASTQ Groomer dataset and click on « Execute »

→ The result of FASTQC is an html page that you can view by clicking on the eye
FASTQC Metrics

- Look at the different metrics for reads
- **Problem**: the per base sequence quality are quite low towards the end

![Per base sequence quality](image)

**Solution:**
Trim the 50bp from the 3’ end of the reads
→ Higher confidence in the sequenced information
FASTQ Trimmer to improve reads quality

1. Use «FASTQ Trimmer» to cut off 50bp from 3’ (use the «search tools» object to find the tool)

2. Run «FASTQC» on the trimmed reads

FASTQ Trimmer (version 1.0.0)

- FASTQ File: 11: FASTQ Groomer on data 2
- Define Base Offsets as: Absolute Values
- Offset from 5’ end: 0
- Offset from 3’ end: 50
- Keep reads with zero length: [ ]
Galaxy Workflow

- **Conversion to Galaxy Format**
  - *Groomer*

- **Reads**
  - (Fastq)

- **Reference Genome**
  - (Fasta)

- **Mapping**
  - *Bowtie2*

- **Quality Control**
  - *FastQC*

- **PCR duplicates Marking**
  - *MarkDup*

- **Mapping statistics With flagstat**

- **Not for PGM data**
Mapping on a reference Genome

• Reads are aligned \( \geq 1 \) times on the reference genome

• A **mapping quality** is associated to each alignment:
  
  → Quantify the probability that the alignment is correct
  
  → Decreases with the number of mismatches (wrong nucleotide) & gaps (small insertions/deletions) & the number of alignments
Reads Alignment - Vocabulary

Mapping method: seed & extend

1. Aligning the seed (small part of the read)
2. Extending the seed to align the whole read

Mismatch: Incoherence between two nucleotides

Indels: Insertion/Deletion into the reference genome

Gap: Bridge within the read alignment (i.e. small indels)

Mappability: Uniqueness of a region

- repeated region = low mappability
- unique region = good mappability
Multiple Alignments

- A read can align **multiple times** on the genome (repeated elements…)
- How to deal with these multiple alignments reads?

Three strategies:

- **1-** Report only unique alignment
- **2-** Report best alignments & randomly assign reads across equally good loci
- **3-** Report all (best) alignments

- **Mapping Quality:** quantify the probability that a read is misplaced
  - Low if a read has multiple alignments ; several mismatches/gaps
Statistics used as Quality Control

- **Depth of coverage** = mean number of reads covering a base (X)
  
  Example: 30X for normal sample, 100X for tumor sample

- **Coverage** = part of the reference with at least one read
  
  Example: >=80% of your exome target is covered by 20X
Alignment tools

• Multitude of alignment tools: BWA, Bowtie, Bowtie2, Bfast…

• How to choose the best tool?
  ▪ Is my sequencing technology supported?
  ▪ Do I have short/long reads? Reads of different sizes?
  ▪ Do I want to allow gaps? Multiple alignments?
  ▪ Does it support single/paired-end reads?
  ▪ On which alignment algorithm is it based?
  ▪ Computational issues?
  ▪ Is it used by the community?
Mapping with Bowtie2

1. Use « Bowtie2 to align reads on the hg19 genome

Bowtie2 (version 0.2)

Is this library mate-paired?:
- Paired-end

FASTQ file:
- 5: FASTQ Groomer on data 3
Nucleotide-space: Must have Sanger-scaled quality values

FASTQ file:
- 9: FASTQ Trimmer on data 6
Nucleotide-space: Must have Sanger-scaled quality values

Minimum insert size for valid paired-end alignments:
- 0

Maximum insert size for valid paired-end alignments:
- 250

Write unaligned reads to separate file(s):
- No

Will you select a reference genome from your history?
- Use a built-in index
Build-ins were indexed using default options

Select a reference genome:
- Homo sapiens hg19

Specify the read group for this file?:
- Yes

Read group identifier (ID). Each @RG I header section:
- chr4
Required if RG specified. Read group ID

Library name (LB):
- chr4
Required if RG specified

Platform/technology used to produce:
- illumina
Required if RG specified. Valid values:

Sample (SM):
- chr4
Required if RG specified. Use pool name

Preset option: combination of parameters designed to have a good tradeoff between speed, sensitivity, accuracy

Execute
SAM/BAM aligned format

- **SAM Format**: aligned format, human readable

```plaintext
@SQ SN:chr12 LN:133851895
@RG ID:Sample_ID LB:Sample_Library PL:ILLUMINA SM:Sample_Name PU:Platform_Unit
```

<table>
<thead>
<tr>
<th>Read name</th>
<th>Flag</th>
<th>Chr</th>
<th>5' pos</th>
<th>MAPQ</th>
<th>Cigar</th>
<th>paired 5' pos of the mate</th>
<th>Insert size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR166338.1</td>
<td>99</td>
<td>chr12</td>
<td>82670685</td>
<td>23</td>
<td>101M</td>
<td>=</td>
<td>82670850 266</td>
</tr>
<tr>
<td>GCCCCTGGGGATTTTTGCACCAAGCCACTGTCTCCAGCTGG</td>
<td>sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBC@GIIHGCFCIEHEAIEIFFGEONDNJFINIONHNGJNNNKNJN</td>
<td>Base quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group affiliation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **BAM Format**: Binary SAM Format (not human readable but compressed = smaller)

Mapping Statistics

- Use « Flagstat » from « Samtools » to see some mapping statistics

flagstat (version 1.0.0)

BAM File to Convert:
8: SAM-to-BAM on data 1 and data 4: converted BAM

Execute

109073 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
107254 + 0 mapped % of mapped reads
0 + 0 paired in sequence
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (–nan%:–nan%)
0 + 0 with itself and mate mapped
0 + 0 singletons (–nan%:–nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
Removing Duplicates

• **Duplicates reads**: different reads having the same sequence caused by PCR amplification during sequencing library preparation

• The removal of the duplicates depends on the application (not suitable for sequencing on small target)

  ![Sequencing error propagated in duplicates](image)

  ![PCRdup removal](image)

• **Galaxy**: Use “Mark Duplicates reads” from “NGS:Picard” to **mark** duplicates (don’t remove them)

  → If duplicates are marked, samtools and GATK tools will ignore them

• **Galaxy**: Run “Flagstat” on the output BAM to see the number of PCR duplicates
Extract workflow from history

• In the « history » panel, click on the topside wheel then
• on « Extract Workflow »
• Write a name for your workflow then click on

![Image of workflow extraction interface]

  • Select the steps you want to see
  • Check every steps!
Select Libraries on Galaxy

You can only run tools on data that are present in your current history

1. In the Data Libraries

2. Select « chr4.fastq »; « chr4.fa »

3. Select « Import to Histories » then click on Go

4. Write a new history name and click on « import library datasets »
Let Galaxy work for you!

Successfully ran workflow ‘Workflow constructed from history ‘TP Initiation’’. The following datasets have been added to the queue:

1. chr4.fe
2. chr4.fastq
22: FASTQ Groomer on data 2
23: FastQC_FASTQ Groomer on data 2.html
24: FASTQ Trimmer on data 22
25: FastQC_FASTQ Trimmer on data 22.html
26: Bowtie2 on data 1 and data 24: aligned reads
27: flagstat on data 26
28: MarkDups_Dupes Marked.bam
29: MarkDups_Dupes Marked.html