Data Preprocessing

Next Generation Sequencing analysis
DTU Bioinformatics
Generalized NGS analysis

Data size

Question
Raw reads
Pre-processing
Assembly: Alignment / de novo
Application specific: Variant calling, count matrix, ...
Compare samples / methods
Answer?
Generalized NGS analysis

Data size

Question | Raw reads | Pre-processing | Assembly: Alignment / de novo | Application specific: Variant calling, count matrix, ... | Compare samples / methods | Answer?

36626 - Next Generation Sequencing Analysis
Assembly: Two basic approaches

- **Alignment**: Use a reference genome and align your reads to the genome

- **de novo assembly**: Try to assemble the reads into a genome without any prior knowledge
Assembly: Two basic approaches

- Alignment: Use a reference genome and align your reads to the genome

- *de novo* assembly: Try to assemble the reads into a genome without any prior knowledge
Assembly: Two basic approaches

- Alignment: Use a reference genome and align your reads to the genome

- de novo assembly: Try to assemble the reads into a genome without any prior knowledge

But first a look at data preprocessing
Preprocessing

• Reads have qualities - bases are not always correct!
• Different error profiles pr. technology
• What can we do?
  • Quality trimming
  • Adaptor clipping
  • 5’ clipping
  • k-mer correction
  • ...
Analyze data using FastQC

- Report basic statistics on your data
- Identify issues with your data

**Basic Statistics**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>tmp.fastq</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Illumina 1.5</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>250000</td>
</tr>
<tr>
<td>Filtered Sequences</td>
<td>0</td>
</tr>
<tr>
<td>Sequence length</td>
<td>101</td>
</tr>
<tr>
<td>%GC</td>
<td>51</td>
</tr>
</tbody>
</table>
Per base sequence quality

Illumina

Quality scores across all bases (Illumina 1.5 encoding)

Trim from 3’ to qual 20
Illumina

Remove reads with avg. qual < 20

Remove reads with “N” basecalls
Trim from 5’

- Sometimes something is fishy in the beginning of the read

Clip a certain number of bases from 5’
Adapters

• Sometimes adapters/primers are also part of the read
• Adapter/primers are non-biological sequences

• Short read alignment is global - adapters are no-go
• de novo assembly will be confused ~ artificial repeats

• If you dont know which were used: FastQC will (may) find them for you!
Adapters - example

We will use “Cutadapt” and “AdapterRemoval” to cut adapters, many other options exist.

Very important if your DNA fragment is shorter than read length.
Main problem is indels at homopolymer runs

(Trim homopolymers), trim trailing poor quality bases

Remove very short reads

For *de novo* adapters should be removed (prinseq)

For alignment we use Smith-Waterman (local) so less important
**k-mer correction**

- What is a k-mer?
- Create a sliding window of size $k$, move it over all your reads and count occurrence of $k$-mers
- We can use this to correct sequencing errors!

**DNA:** ACGTGTAAACGTGACGTTGGA

**Eg. $k=5$**

ACGTG
CGTGT
GTGTA
**k-mer correction**

**Concept:** Rare k-mers are seq. errors
Need >15X coverage

- ACGTGGTTGCCCTTTAAA
- ACGTGGTTAACCCTTTAAA
- ACGTGGTTAACCCTTTAAA
- ACGTGGTTAACCCTTTAAA
- ACGTGGTTAACCCTTTAAA
- ACGTGGTTAACCCTTTAAA
- ACGTGGTTAACCTTTAAA
- ACGTGGTTAACCTTTAAA
- ACGTGGTTAACCTTTAAA
- ACGTGGTTAACCTTTAAA

*Kelley et al., 2010*
Merge paired ends

- Insert size: 500nt
- Reads: 100nt
- Middle: 300nt

- Insert size: 180nt
- Reads: 100nt
- Middle: -20nt

- Merge overlapping pairs: single longer read
- Smart because Illumina reads have bad 3’ quals
- Very useful for de novo assembly

Magocˇ and Salzberg, 2011
Merge paired ends

- Merge overlapping pairs: single longer read
- Smart because Illumina reads have bad 3' quals
- Very useful for de novo assembly

Magoc and Salzberg, 2011
Coverage

- Coverage/depth is how many times that your data covers the genome (on average)

- Example:
  - N: Number of reads: 5 mill
  - L: Read length: 100
  - G: Genome size: 5 Mbases
  - \( C = \frac{N \times L}{G} \)
  - On average there are 100 reads covering each position in the genome
Last, but important!

• Lots of data - storage is expensive!

• Keep data compressed whenever possible (gzip, bzip, bam)

• Remove intermediate files and files that can easily be re-created