On the homework

- Being unable to improve the quality of the data means one of the two:
  1. I gave up too easily
  2. I would be removing too much data

*Every dataset can be improved dramatically!* At the cost of losing data (measurements) and potential for introducing artifacts.

Reminder

- Periodically re-check the documentation for small but essential details.

Example:

- **bwa** needs to be indexed differently for small and large genomes
- **bwa** has to be invoked with different alignment modes for **short reads** (200 <) and **long reads** (200 >)
More on SAM/BAM formats

• SAM/BAM are the starting point of most analyses

• You can answer a surprisingly large number of questions with just samtools

Sequencing Coverage (Depth)

Lander/Waterman model

1. random reads
2. ability to detect overlap does not change

coverage \( C = N \times L / G \)

\( N \) = number of reads, \( L \) = length of reads, \( G \) = size of genome

Probability of a base not being sequenced

\( P = \exp(-C) \)

To get the percent of genome not covered (multiply by 100)

\( N=35 \text{ million}, L=35, G=250 \text{ million} \)

\( C = 5 \rightarrow 0.6\% \text{ genome not sequenced} \rightarrow 15 \text{ million bases not covered} \)

Realistic coverage measures

Neither of the models assumptions are correct

→ empirically → raise the required coverage at least 10 fold

What part of the genome is coverable to begin with?

Also known as “accessible”, “mappable” portion

→ 90% human genome

Paired end sequencing

• Nomenclature: paired-end and mated-pairs are different technologies

• The net effect is obtaining one other piece of information: ability to connect to distant reads as coming from one pair.
Paired end (PE) sequencing (most common)

Sequences both end of the same DNA fragment

We end up with two reads that are known to have come from the different strands of the same DNA fragment – insert sizes 200-600bp

Notation:

Mated-pair (MP) sequencing

DNA is circularized then cut,

SOLiD Mate-Pair protocol

Notation:

Dealing with paired data

• Make sure to understand which parts of the DNA fragments have been sequenced. Consult your sequencing operator for details on the library preparation.

• When in doubt you can always map them in single end mode, then visualize the results (covered in later lectures) and see how the pairs are located relative to one another. (sanity check)

• Consult vendor materials → comprehensive but will also contain a lot of details that are not relevant – not easy to make sense of these

More strategies

• Just about all aligners can deal with standard paired end (PE) sequencing data

• A few can deal with mate-pair (MP) and their variations → see Novoalign, check vendor recommended tools

• Finally you may turn the pairs into standard PE by reverse complementing the proper reads.
Quick checklist

• How are my pairs oriented?

• How is the data formatted?
  – are the reads in the same file (interleaved?)
  – are the reads in separate files?
  – what is the naming convention?
  – what is the insert size and its size distribution (minimum, maximum insert sizes)

Summary: paired end vs mated pairs

• Paired ends is supported by some technologies where it is possible to sequence from both ends of a clone.

• Mate pairs involves making circular fragments using a linker sequence, and fragmenting them around the linker, and then sequencing the result.

• The distance between mate pairs are much longer (2-5kb), while paired-end fragments are rarely more than 500bp apart.

• The technologies keep evolving within a year → make sure to ask questions from the facility managers!

Read simulator in the samtools package

Generating sequencing data with known properties – then try to detect the known features – some generators include error models for sequencing platforms

```
ialbert@porhos ~
$ cd bin
ialbert@porhos ~/bin
$ ln -s /src/samtools-0.1.18/misc/wgsim wgsim
ialbert@porhos ~/bin
$ cd
ialbert@porhos ~
$ ~/bin/wgsim
```

wgsim

```
Usage: wgsim [options] <in.ref.fai> <out.read1.fq> <out.read2.fq>
Options: -e FLTERT base error rate [0.020]
        -d INT outer distance between the two ends [500]
        -s INT standard deviation [50]
        -n INT number of read pairs [1000000]
        -l INT length of the first read [70]
        -z INT length of the second read [70]
        -x FLTERT rate of mutations [0.0001]
        -n FLTERT fraction of indels [0.15]
        -s FLTERT probability an indel is extended [0.30]
        -d INT seed for random generator [-1]
        -h haplotype mode
```
Internally a two step process

1. Generate a mutated genome.

2. Select reads from this mutated genome

Simulate paired end reads

<table>
<thead>
<tr>
<th>Reference</th>
<th>Mutation</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Heterozygous mutation of T into S (G or C)
Homoyzgous mutation of C into G
Insertion of A into reference
Deletion of C from reference

Generate error free reads from an unmodified genome

Paired end alignments

These two lines change the rest is the same
Filtering your paired alignment

Explain SAM flags: http://picard.sourceforge.net/explain-flags.html

### Homework 13

Show the commands and the answers for the following:

- Generate **100000 read pairs** from the yeast genome with a base error rate of **0.01**

- Generate paired end indexed **BAM** file from the data.

- How many of your reads are mapped in **proper pair**?

- How many of your **unique reads** have the **first of the pair** on the **forward strand**?