Reminder

• Before any serious work 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 >)
Sequencing Coverage (Depth)

Lander/Waterman model: 1. random reads 2. ability to detect overlap does not change

coverage \rightarrow 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 million, L=35, G=250 million \rightarrow C = 5 \rightarrow 0.6\% genome not sequenced \rightarrow 15 million bases not covered
Realistic coverage

Neither of the models assumptions are correct
→ multiply 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 (PE) sequencing

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: insert
Mated-pair (MP) sequencing

DNA is circularized then cut,

SOLiD Mate-Pair protocol

Notation:

- reverse strand
- forward strand

insert size

- Same strand
- R3
- F3

Notation: insert

mated pair insert sizes → 2000 – 5000bp long

(may change as new protocols are developed)
Dealing with paired data

• Make sure to understand where the results come from. 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.
Read simulators

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

```
$ ~/bin/wgsim

Program: wgsim (short read simulator)  
Version: 0.2.3  
Contact: Heng Li <lh3@sanger.ac.uk>

Usage:  wgsim [options]  <in.ref.fa>  <out.read1.fq>  <out.read2.fq>

Options:  
-e FLOAT  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]  
-1 INT  length of the first read [70]  
-2 INT  length of the second read [70]  
-r FLOAT  rate of mutations [0.0010]  
-R FLOAT  fraction of indels [0.10]  
-X FLOAT  probability an indel is extended [0.30]  
-c  generate reads in color space (SOLiD reads)  
-C  show mismatch info in comment rather than read n  
-h  haplotype mode
```
Simulate paired end reads

```
$ wgsim -N 250000 ~/refs/yeast.fasta r1.fq r2.fq > out
```

```
[wgsim_core] 18 sequences, total length: 12163423
```

```
# the mutation output lists the changes inserted into the reads

$ head -5 out
chr01 1498  T   Y   +
chr01 1568  G   S   +
chr01 2672  A   C   -
chr01 3349  T   Y   +
chr01 3846  C   Y   +
```
Align paired end reads with BWA

```
# the paired reads are in separate file

# align reads from file 1
bwa aln ~/refs/yeast.fasta r1.fq > r1.aln

# align reads from file 2
bwa aln ~/refs/yeast.fasta r2.fq > r2.aln

# generate paired end SAM file
bwa sampe ~/refs/yeast.fasta r1.aln r2.aln r1.fq r2.fq > out.sam
```

Above we are using the default insert sizes, look at the command options to see how to set these. We also create an indexed bam file from the out.sam file (see lecture 7)
Filtering your paired alignment

Explain SAM flags: http://picard.sourceforge.net/explain-flags.html
Homework 8

Show the commands for:

• Generate 250000 read pairs from the yeast genome with an insert size of 1kb

• Align these reads with bwa then generate an paired end indexed BAM file

• 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?

  (hint you may need to use all of –q, –f and –F flags)