Paired end sequencing

- More information: the instrument is able to connect reads that belong to the original fragment
- Nomenclature: paired-end and mated-pairs are different technologies
- The technology is vendor specific with quirks and tacit assumptions

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

File 1

File 2

Note: pairing may imply orientation but only if the sequencing protocol was also strand specific

Paired end (PE) sequencing

short fragments, long reads

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

Read merging/stiching
Mated-pair (MP) sequencing

SOLID Mate-Pair protocol

- Same strand
- F3
- R3

mated pair insert sizes → 2000 – 5000bp long
(may change as new protocols are developed)

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 operate in single end mode, then visualize the results (covered in later lectures)
- Verify how the pairs are located relative to one another. (sanity check, see later)
- Consult vendor materials → comprehensive but will also contain a lot of details that are not relevant

Back to the FastQC reports

Tips:
1. Understand the plots that you see
2. Don’t just look at the first plot and move on
3. Methods that rely on counting reads RNA-Seq, Chip-Seq → sensitive to duplication rates
4. Methods that rely on assembling unknown genomes/transcriptomes → sensitive to base calling errors
5. It is possible to overcorrect! There may be systematic reasons why some genomic regions are not covered well. If you just cut those out you may alter your data in unanticipated ways.

Tools used in the ebola genome paper.

- Quality control: Picard, FastQC, Trimmomatic, trimAl, PRINSEQ, BMTagger
- Metagenomic Screen: RDP3 (suite of techniques)
- Alignment: Lastal, Novoalign, MAFFT, MUSCLE
- SNP calling: GATK, SNPEff, Blast
- Assembly: Trinity, MetaVelvet
- Phylogeny: RAxML, MrBayes, Path-O-Gen
- Molecular dating: BEAST, TreeAnnotator
- Visualization: Geneious, MEGAN

That is a lot of tools/methods. Ask yourself why do we need so many?
Quality Trimming: before \(\rightarrow\) after

Sequence length distributions before-after

Automate repetitive tasks with simple scripts

Homework 8

- Select three datasets from the ebola sequencing runs.
- Run a base quality filtering step with the tool of your choice.
- Run fastqc reports on this data.
- Provide a short discussion (one paragraph) on how the data changed by the process of base quality filtering.
- Does the data seem to have the same properties before/after quality filtering?