Week 6, Lecture 12

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center

Penn State
Install bioawk

- **bioawk** - a version of awk that is bioinformatics aware

- Can read **fastq** files one record at a time

- Can reverse complement sequences, understands many bioinformatics file formats
bioawk examples

See more details and examples on the course webpage
Sequencing technologies

- DNA → library prep (fragments) → sequencing
- DNA is double stranded
- Sequencing operates on a single strand thus each strand will be sequenced separately
Paired end sequencing

• More information: 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
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)

- Consult vendor materials comprehensive but will also contain a lot of details that are not relevant
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.
Competing representations

SE – single end reads, PE – paired end reads

Paired end reads come in either

• two files with the exact same number of lines and IDs, where a pair is present on the same line”

• a single file where pairs are consecutive records (interleaved)
The read order is now also essential

Regardless of representation one now needs to ensure that the order of reads will keep matching

Read removal needs to take place on both files or both lines if the file is interleaved.
Quick PE 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 expected insert (fragment) size and its 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!
optional: install Trimmomatic

• It is a great tool to deal with **paired end reads**

• Lacks some options that **cutadapt** has

• But it has options **cutadapt** does not directly support
Homework 12

• Download dataset \texttt{lect12.tar.gz}

• It contains a paired end read dataset

• Using techniques learned in the past two lectures identify problems that the data may have and improve its overall quality