Read alignment

Learning outcomes

After having completed this chapter you will be able to:

- Explain what a sequence aligner does
- Explain why in some cases the aligner needs to be 'splice-aware'
- Calculate mapping quality out of the probability that a mapping position is wrong
- Build an index of the reference and perform an alignment of paired-end reads with `bowtie2`

Material

Download the presentation

- Unix command line E-utilities documentation
- `bowtie2` manual
- Ben Langmead's youtube channel for excellent lectures on e.g. BWT, suffix matrixes/trees, and binary search.

Exercises

Prepare the reference sequence

Make a script called `05_download_ecoli_reference.sh`, and paste in the code snippet below. Use it to retrieve the reference sequence using `esearch` and `efetch`:

05_download_ecoli_reference.sh

```
#!/usr/bin/env bash

REFERENCE_DIR=~/project/ref_genome/

mkdir $REFERENCE_DIR
cd $REFERENCE_DIR

esearch -db nuccore -query 'U00096' \
| efetch -format fasta > ecoli-strK12-MG1655.fasta
```

Exercise: Check out the documentation of `bowtie2-build`, and build the indexed reference genome with bowtie2 using default options. Do that with a script called `06_build_bowtie_index.sh`.

Answer

06_build_bowtie_index.sh

```
#!/usr/bin/env bash

cd ~/project/ref_genome

bowtie2-build ecoli-strK12-MG1655.fasta ecoli-strK12-MG1655.fasta
```

Align the reads with `bowtie2`

Exercise: Check out the bowtie2 manual here. We are going to align the sequences in paired-end mode. What are the options we'll minimally need?

Answer

According to the usage of `bowtie2`:

```
bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i> |
--sra-acc <acc> | b <bam>}
```

We'll need the options:

- `-x` to point to our index
- `-1` and `-2` to point to our forward and reverse reads

Exercise: Try to understand what the script below does. After that copy it to a script called `07_align_reads.sh`, and run it.

07_align_reads.sh

```
#!/usr/bin/env bash

TRIMMED_DIR=~/project/results/trimmed
REFERENCE_DIR=~/project/ref_genome/
ALIGNED_DIR=~/project/results/alignments

mkdir -p $ALIGNED_DIR

bowtie2 \
-x $REFERENCE_DIR/ecoli-strK12-MG1655.fasta \
-1 $TRIMMED_DIR/trimmed_SRR519926_1.fastq \
-2 $TRIMMED_DIR/trimmed_SRR519926_2.fastq \
> $ALIGNED_DIR/SRR519926.sam
```

We'll go deeper into alignment statistics later on, but `bowtie2` writes already some statistics to stdout. General alignment rates seem okay, but there are quite some non-concordant alignments. That doesn't sound good. Check out the explanation about concordance at the bowtie2 manual. Can you guess what the reason could be?