Mapping with BWA

Objectives

In this lab, you will explore a popular fast mapper called BWA. Simulated RNA-seq data will be provided to you; the data contains 75 bp paired-end reads that have been generated in silico to replicate real gene count data from Drosophila. The data simulates two biological groups with three biological replicates per group (6 samples total). The objectives of this lab is mainly to:

1. Learn how BWA works and how to use it.

Introduction

BWA (the Burrows-Wheeler Aligner) is a fast short read aligner. It is an unspliced mapper. It's the successor to another aligner you might have used or heard of called MAQ (Mapping and Assembly with Quality). As the name suggests, it uses the burrows-wheeler transform to perform alignment in a time and memory efficient manner.

BWA Variants

BWA has three different algorithms:

- For reads upto 100 bp long:
  - BWA-backtrack : BWA aln/samse/sampe
- For reads upto 1 Mbp long:
  - BWA-SW
  - **BWA-MEM** : Newer! Typically faster and more accurate.

Get your data

Six raw data files have been provided for all our further RNA-seq analysis:

- c1_r1, c1_r2, c1_r3 from the first biological condition
- c2_r1, c2_r2, and c2_r3 from the second biological condition

Get set up for the exercises

```
cds
cd my_rnaseq_course
cd day_2/bwa_exercise
```

Lets look at the data files and reference files

Get set up for the exercises

```
ls ../data
ls ../reference

#transcriptome
head ../reference/transcripts.fasta
#see how many transcripts there are in the file
grep -c '^>' ../reference/transcripts.fasta

#genome
head ../reference/genome.fa
#see how many sequences there are in the file
grep -c '^>' ../reference/genome.fa


#annotation
head ../reference/genes.formatted.gtf
#see how many entries there are in this file
wc -l ../reference/genes.formatted.gtf
```

Run BWA

Load the module:

```
module load intel/17.0.4
module load bwa
```

You can see the different commands available under the bwa package from the command line help:

```
bwa
```

Part 1. Create a index of your reference

NO NEED TO RUN THIS NOW- YOUR INDEX HAS ALREADY BEEN BUILT!

All the files starting with the prefix transcripts.fasta are your BWA index files.

```
bwa index -a bwtsw reference/transcripts.fasta
```

Part 2. Align the samples to reference using bwa mem

Running alignment using the newest and greatest, BWA MEM to the transcriptome. Alignment is just one single step with bwa mem.

Submit to the TACC queue or run in idev session

Create a `commands` file and use ***launcher_creator.py*** followed by ***sbatch***.

Make sure each command is one line in your commands file.

Put this in your commands file

```
nano commands.mem

#Enter these lines into the file
bwa mem -o C1_R1.mem.sam ../reference/transcripts.fasta ../data/GSM794483_C1_R1_1.fq ../data
/GSM794483_C1_R1_2.fq
bwa mem -o C1_R2.mem.sam ../reference/transcripts.fasta ../data/GSM794484_C1_R2_1.fq ../data
/GSM794484_C1_R2_2.fq
bwa mem -o C1_R3.mem.sam ../reference/transcripts.fasta ../data/GSM794485_C1_R3_1.fq ../data
/GSM794485_C1_R3_2.fq
bwa mem -o C2_R1.mem.sam ../reference/transcripts.fasta ../data/GSM794486_C2_R1_1.fq ../data
/GSM794486_C2_R1_2.fq
bwa mem -o C2_R2.mem.sam ../reference/transcripts.fasta ../data/GSM794487_C2_R2_1.fq ../data
/GSM794487_C2_R2_2.fq
bwa mem -o C2_R3.mem.sam ../reference/transcripts.fasta ../data/GSM794488_C2_R3_1.fq ../data
/GSM794488_C2_R3_2.fq
```

launcher_creator.py -n mem -t 04:00:00 -j commands.mem -q normal -a UT-2015-05-18 -m "module load intel/17.0.4;module load bwa" -l bwa_mem_launcher.slurm
sbatch --reservation=RNASeq-Tue bwa_mem_launcher.slurm

#or if reservation is giving us issues

sbatch bwa_mem_launcher.slurm

Since this will take a while to run, you can look at already generated results at: bwa_mem_results_transcriptome

Alternatively, we can also use bwa to map to the genome (reference/genome.fa).

Now that we are done mapping, lets look at how to assess mapping results.