RNA-seq mapping practical

Ernest Turro
Romina Petersen
Luigi Grassi
University of Cambridge

14 September 2016

1 Introduction

In this practical we shall map RNA-seq reads from a study of the *ps* splice factor in *Drosophila melanogaster* cell cultures [1]. The dataset consists of a treatment and a control group. The treatment group comprises three cell cultures in which the *pasilla* splice factor has been knocked down. The remaining four cell cultures are untreated and serve as a control.

At each step, please pay careful attention to the commands before you run them, making sure you understand what they do and why.

2 Preliminaries

The practical employs or refers to the following software:

- Integrative Genomics Browser version $\geq$ 2.3.40 (`http://www.broadinstitute.org/igv`)
- SAMtools version 0.1.19 (`http://www.htslib.org`)
- FASTX toolkit version 0.0.13 (`http://hannonlab.cshl.edu/fastx_toolkit`)
- Bowtie aligner version 1.1.1 (`http://bowtie-bio.sf.net`)
- TopHat gapped aligner version 1.4.1 (`http://tophat.cbcb.umd.edu`)
- kallisto version 0.42.5 (`https://pachterlab.github.io/kallisto`)

3 Servers

Most of this practical will be conducted on a remote server. To log in, follow these instructions:

1. Go to the course server web page using a web browser (e.g. Google Chrome).
2. Click on "Terminal/ssh" press enters/returns for IPs/host and port (defaults are fine).
3. Enter login credentials.

4 Gapped genome alignment

Alignment is a computationally demanding task. It is therefore unusual to attempt to perform alignment on a realistic dataset during a practical. However, today, we will use a small fastq file.

4.1 The Bowtie1 index

The Bowtie1 index takes the form of a collection of files ending in `.ebwt` which encode a compact and structured representation of FASTA sequences. The index can be used by the Bowtie1 and TopHat1 aligners to map short reads to the reference sequences.

Open a session on the server and change directory (`` `cd` ``) to the `/data/day3/practical/ref` directory. The genome FASTA and Bowtie1 files are contained in that directory with the prefix `Dmel.BDGP5`. The index contains a subset of the chromosomes in the Ensembl file `Drosophila_melanogaster.BDGP5.68.dna.toplevel.fa` (excludes the heterochromatic chromosomes).

- How many chromosomes does the *D. melanogaster* euchromatic genome have? (hint: use the `grep` command).
- You can download a reference genome from several websites as NCBI, ENSEMBL or UCSC (`http://www.ncbi.nlm.nih.gov/genome/`, `http://www.ensembl.org/info/data/ftp/index.html`, `http://hgdownload.cse.ucsc.edu/downloads.html`).
- What command was used to generate the `Dmel.BDGP5.*.ebwt` files? (hint: check the Bowtie1 manual).

4.2 The reads

The reads are stored in gzipped FASTQ files which were downloaded from the European Nucleotide Archive (`http://www.ebi.ac.uk/ena`) and placed in subfolders within the `~/data/day3/practical/all_reads` folder.

Note that to extract text from gzipped files you need to use the `zcat` command.

- How many FASTQ files are there for the accession ID GSM461179?
- Are the reads single or paired-end?
- Try locating these reads on the ENA web site — how many megabases are there in total for GSM461179?

4.3 Trimming

The last bases of the reads in this dataset tend to be of poor quality. You could see this using the FastQC program or you could just take a peek at some of the FASTQ files by eye:

- Try running the `head` command on `GSM461176_untreated1/SRR031728.fastq.gz` and pipe to gunzip in the `all_reads` directory — how can you tell that there is a problem with the base qualities of the last bases of the reads? (hint: visit `http://en.wikipedia.org/wiki/FASTQ_format#Encoding`)

All the FASTQ files have been trimmed down to 37bp except for `SRR031718_00_untrimmed.fastq.gz`. Load the `fastx_toolkit` module and go in the `~/scratch/day3` directory, then try running `fastx_trimmer` to trim that last file down. Note that the '\' character simply tells the shell to continue the command on the next line and should be omitted if you enter the entire command in a single line.

```
module load fastx_toolkit/0.0.13.2
cd ~/scratch/day3
fastx_trimmer -h # run this to see the documentation
zcat /data/day3/practical/split_reads/GSM461176_untreated1/\
SRR031728_00_untrimmed.fastq.gz\
| fastx_trimmer  -f 1 -l 37 -Q33 | gzip -c > SRR031728_00.fastq.gz
```

Peek into the new file using `head` and make sure the trimming was successful.

- Can you work out why the `-Q33` option is necessary?

4.4 TopHat1 alignment

At this stage you are going to align two sets of reads — one single and one paired-end. First open two different terminals (you can open multiple terminals by clicking the plus symbol on the right side of the window) and change directory to `~/scratch/day3` in each one. We will be running several instances of TopHat1 simultaneously. First familiarise yourself with the TopHat1 manual (`tophat -h`). Note that there should be *no space character* after the comma separating file names. Note also that the '\' character simply tells the shell to continue the command on the next line and should be omitted if you enter the entire command in a single line. The commands may take a while to complete.

1. `module load tophat/1.4.1`
2. `cd /data/day3/practical/split_reads/GSM461176_untreated1/`

   `tophat --segment-length 18 -o ~/scratch/day3/GSM461176_untreated1 \`

   `Dmel.BDGP5 ~/scratch/day3/SRR031728_00.fastq.gz,SRR031729_00.fastq.gz`
3. `cd /data/day3/practical/split_reads/GSM461178_untreated3/`

   `tophat --segment-length 18 -r 120 -o ~/scratch/day3/GSM461178_untreated3 \`

   `Dmel.BDGP5 SRR031714_1_00.fastq.gz,SRR031715_1_00.fastq.gz \`

   `SRR031714_2_00.fastq.gz,SRR031715_2_00.fastq.gz`

- What does the `-r` option do and what is its relation to the insert size and the fragment size?
- Why is the `--segment-length` parameter set to 18?
- How does `tophat` know where to find the Bowtie1 index?

Once the alignments have been completed give a look to the results:

- Where are the alignments stored?
- Print out the headers for the aligned read (BAM) files using `samtools`
- What are the alignment rates?

5 Visualising alignments

Note: any files required for this section can be downloaded using the Apache server links on the course server.

Launch the Integrative Genomics Browser (IGV). Download the Dmel.BDGP5.fa and Drosophila_melanogaster.BDGP5.25.68.gtf from the Apache server links on the course server (in the folder /data/day3/practical/ref/) and accepted_hits.bam (from the folder /scratch/day3/GSM461176_untreated1)

```
Genomes --> Load Genome --> Dmel.BDGP5.fa
```

Now load the gene annotations, which are stored as a GTF file (don't bother indexing if prompted by IGV):

```
File --> Load from File --> Drosophila_melanogaster.BDGP5.25.68.gtf
```

Finally, let us load the BAM file for the first sample:

```
File --> Load from File --> accepted_hits.bam
```

As you will see, loading the BAM file will fail because it has not been indexed. Indexing is necessary for fast access to the alignment information. Run `samtools index` on all BAM files created. E.g.:

```
samtools index ~/scratch/day3/GSM461176_untreated1/accepted_hits.bam
```

Then try again to load the BAM file for the first sample into IGV upon you have downloaded the corresponding index from the Apache server links on the course server . Have a look around the first 20kb of the 2L chromosome.Pay particular attention to the spliced reads and try to get a rough idea of the different isoform structures for gene FBgn0002121 that may be present in the sample. Do you think the alignment coverage is sufficient to evaluate the expression level of the different isoforms?

Download the alignment of the complete set of reads and the corresponding index file (folder /data/day3/practical/th_out/untreated1) and look at the same region again.

6 Ungapped transcriptome alignment or pseudoalignment

We shall now align or pseudoalign a subset of our reads to the transcriptome rather than the genome. Open a terminal on the server and `cd` to the `~/share/Day1/data/mapping/ref` directory. The transcriptome FASTA and the Bowtie1 and kallisto index files are contained in that directory with the prefix `Dmel.BDGP5-transcripts`. The sequences were obtained by merging the cDNA and the non-coding RNA FASTAs from Ensembl.

- Try to locate these files on the Ensembl FTP server
- How many transcripts are there?
- How was the kallisto index constructed?

6.1 Bowtie1 alignment

We shall now align one of the paired-end read files to the full set of transcript sequences:

```
cd ~/scratch/day3
bowtie -a --best --strata -S -m 100 -X 400 --chunkmbs 256 --fullref /data/day3/practical/ref/Dmel.BDGP5-transcripts \

-1 <(gzip -dc /data/day3/practical/all_reads/GSM461178_untreated3/SRR031714_1.fastq.gz) \

-2 <(gzip -dc /data/day3/practical/all_reads/GSM461178_untreated3/SRR031714_2.fastq.gz) | \

samtools view -F 0xC -bS - | \

samtools sort -n - ~/scratch/day3/untreated3_transcriptome
```

While this command runs, take a look at the Bowtie1 documentation and try to work out the function of each of the parameter options. In particular,

- Why might the `-a` flag be important?
- What is the effect of using the `--fullref` option and what additional information might that give us?

Also try to understand the piping to the `samtools` program:

- What does the `-F 0xC` samtools option do? Why might it be a good idea to use it?
- Why might it be useful to sort the reads as in the above command?

Finally, take a look at some of the alignments to gene FBgn0002121:

```
samtools view ~/scratch/day3/untreated3_transcriptome.bam | grep FBgn0002121 | head
```

Pick one or two read pairs and check that the alignment between the transcriptome and the genome BAM files are consistent with each other.

6.2 kallisto pseudoalignment

We shall now run the above mapping procedure using kallisto instead of Bowtie1:

```
cd ~/scratch/day3
kallisto quant --pseudobam -i /data/day3/practical/ref/Dmel.BDGP5-transcripts.kind \

/data/day3/practical/all_reads/GSM461178_untreated3/SRR031714_1.fastq.gz \

-o ~/scratch/day3/untreated3-transcriptome-k /data/day3/practical/all_reads/GSM461178_untreated3/\

SRR031714_2.fastq.gz | \

samtools view -bS - > ~/scratch/day3/untreated3-transcriptome-k.bam
```

- How do you think the fragment length distribution estimated?
- What is the purpose of the `--pseudobam` option?
- How about `--bootstrap-samples`, which we didn't use here?
- Identify and characterise the output of the run

References

[1] Brooks, Angela N., et al. "Conservation of an RNA regulatory map between Drosophila and mammals." Genome research 21.2 (2011): 193-202.