RNA-seq mapping practical

Ernest Turro
University of Cambridge

21 Oct 2013

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 is composed of 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:

- R version 2.15 (`http://www.r-project.org`)
- Integrative Genomics Browser version 2.1.24 (`http://www.broadinstitute.org/igv`)
- SAMtools version 0.1.18 (`http://samtools.sf.net`)
- FASTX toolkit version 0.0.13 (`http://hannonlab.cshl.edu/fastx_toolkit`)
- Bowtie aligner version 0.12.8 (`http://bowtie-bio.sf.net`)
- TopHat gapped aligner version 1.4.1 (`http://tophat.cbcb.umd.edu`)

3 Gapped genome alignment

Alignment is a computationally demanding and time-intensive task. It is therefore very unusual to attempt to perform alignment on a realistic dataset during a practical. However, today, we will attempt to do this in a distributed fashion across all 40 computers. Each of you will be given an integer N between 0 and 39 and will be responsible for aligning 1/40th of the reads or read pairs in the *pasilla* dataset. Once everyone has aligned their chunk, you will merge your alignments to obtain a complete set of alignments each. Please make sure it absolutely clear to you what your value of N is and that it is different from the value for other participants before you start!

3.1 The Bowtie index

The Bowtie index is a collection of files ending in `.ebwt` which contain a compact and structured representation of FASTA sequences. The index can be used by the Bowtie and TopHat aligners to map short reads to the reference sequences.

Open a terminal and change directory (`` `cd` ``) to the `/nfs/training/ref` directory. The genome FASTA and Bowtie 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.BDGP 5.68.dna.toplevel.fa` (basically, excludes the heterochromatic chromosomes).

- How many chromosomes does the *D. melanogaster* euchromatic genome have? (hint: use the `grep` command).
- Can you find the "toplevel" FASTA file on the Ensembl FTP server? (hint: try connecting to the server using `ftp ftp.ensembl.org` with username `anonymous`).
- What command was used to generate the `Dmel.BDGP5.*.ebwt` files? (hint: check the Bowtie manual).

3.2 The reads

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

- 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 — approximately how many megabases are there in total for GSM461179?

If you now change directory to the `/nfs/training/split_reads` directory, you will find that the FASTQ files have been split into 40 chunks labelled 0 to 39. You will be processing the chunks that correspond to your value of N.

3.3 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 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_N.fastq-untrimmed, where N is your unique integer ID. Try running fastx_trimmer to trim that last file down:

```
fastx_trimmer -h # run this to see the documentation
cd /nfs/training/split_reads/GSM461176_untreated1
fastx_trimmer  -f 1 -l 37 -Q33 -i SRR031728_N.fastq-untrimmed -o SRR031728_N.fastq
```

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

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

3.4 TopHat alignment

At this stage you are going to align seven sets of reads — one set for each condition. First open four different terminals (e.g. in different tabs) and change directory to /nfs/training/split_reads in each one. We will be running several instances of TopHat simultaneously. First familiarise yourself with the TopHat manual (tophat -h). The first round of simultaneous commands will align all the reads in the untreated samples. Try running each of these commands in a separate tab, remembering to replace the N with your integer ID using two digits (e.g. use 03 instead of 3). They will take a while to complete:

1.
   ```
   tophat --segment-length 18 -o /nfs/training/tophat_out/untreated1/N \
   Dmel.BDGP5 \
   GSM461176_untreated1/SRR031728_N.fastq,GSM461176_untreated1/SRR031729_N.fastq
   ```

2.
   ```
   tophat --segment-length 18 -o /nfs/training/tophat_out/untreated2/N \
   Dmel.BDGP5 \
   GSM461177_untreated2/SRR031708_N.fastq,GSM461177_untreated2/SRR031709_N.fastq,\
   GSM461177_untreated2/SRR031710_N.fastq,GSM461177_untreated2/SRR031711_N.fastq,\
   GSM461177_untreated2/SRR031712_N.fastq,GSM461177_untreated2/SRR031713_N.fastq
   ```

3.
   ```
   tophat --segment-length 18 -r 120 -o /nfs/training/tophat_out/untreated3/N \
   Dmel.BDGP5 \
   GSM461178_untreated3/SRR031714_1_N.fastq,GSM461178_untreated3/SRR031715_1_N.fastq \
   GSM461178_untreated3/SRR031714_2_N.fastq,GSM461178_untreated3/SRR031715_2_N.fastq
   ```

4.
   ```
   tophat --segment-length 18 -r 120 -o /nfs/training/tophat_out/untreated4/N \
   Dmel.BDGP5 \
   GSM461182_untreated4/SRR031716_1_N.fastq,GSM461182_untreated4/SRR031717_1_N.fastq \
   GSM461182_untreated4/SRR031716_2_N.fastq,GSM461182_untreated4/SRR031717_2_N.fastq
   ```

- 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 Bowtie index?

Once the above four commands have completed, align the reads for the treated samples:

1.
   ```
   tophat --segment-length 18 -o /nfs/training/tophat_out/treated1/N \
   Dmel.BDGP5 \
   GSM461179_treated1/SRR031718_N.fastq,GSM461179_treated1/SRR031719_N.fastq,\
   GSM461179_treated1/SRR031720_N.fastq,GSM461179_treated1/SRR031721_N.fastq,\
   GSM461179_treated1/SRR031722_N.fastq,GSM461179_treated1/SRR031723_N.fastq
   ```

2.
   ```
   tophat --segment-length 18 -r 120 -o /nfs/training/tophat_out/treated2/N \
   Dmel.BDGP5 \
   GSM461180_treated2/SRR031724_1_N.fastq,GSM461180_treated2/SRR031725_1_N.fastq \
   GSM461180_treated2/SRR031724_2_N.fastq,GSM461180_treated2/SRR031725_2_N.fastq
   ```

3.
   ```
   tophat --segment-length 18 -r 120 -o /nfs/training/tophat_out/treated3/N \
   Dmel.BDGP5 \
   GSM461181_treated3/SRR031726_1_N.fastq,GSM461181_treated3/SRR031727_1_N.fastq \
   GSM461181_treated3/SRR031726_2_N.fastq,GSM461181_treated3/SRR031727_2_N.fastq
   ```

While the above three commands run in three separate terminal tabs, browse through the `tophat_out` directories in the fourth tab.

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

3.5 Merging BAM files

Once all 40 participants have finished aligning their respective FASTQ files, we shall merge the BAM files using `samtools` to produce a single complete BAM file per sample on each machine. It is *crucial* that you ensure that *everyone* in the class has reached this point in the tutorial before proceeding. Once the instructor has given you the go-ahead, you may run the following commands simultaneously in separate terminal tabs:

-
  ```
  UNTREATED1=(`find /nfs/training/tophat_out/untreated1 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/untreated1.bam ${UNTREATED1[@]}
  ```

-
  ```
  UNTREATED2=(`find /nfs/training/tophat_out/untreated2 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/untreated2.bam ${UNTREATED2[@]}
  ```

-
  ```
  UNTREATED3=(`find /nfs/training/tophat_out/untreated3 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/untreated3.bam ${UNTREATED3[@]}
  ```

-
  ```
  UNTREATED4=(`find /nfs/training/tophat_out/untreated4 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/untreated4.bam ${UNTREATED4[@]}
  ```

And now for the treated samples — you may run these commands simultaneously:

-
  ```
  TREATED1=(`find /nfs/training/tophat_out/treated1 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/treated1.bam ${TREATED1[@]}
  ```

-
  ```
  TREATED2=(`find /nfs/training/tophat_out/treated2 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/treated2.bam ${TREATED2[@]}
  ```

-
  ```
  TREATED3=(`find /nfs/training/tophat_out/treated3 -name accepted_hits.bam`)
  samtools merge -f ~/Desktop/treated3.bam ${TREATED3[@]}
  ```

4 Visualising alignments

Launch the Integrative Genomics Browser (IGV) (run `igv.sh &` from the command line) and load the genome FASTA file:

```
File --> Load Genome --> select /nfs/training/ref/Dmel.BDGP5.fa
```

Now load the gene annotations, which are stored as a GTF file:

```
File --> Load from File --> select /nfs/training/ref/Drosophila_melanogaster.BDGP5.25.68.gtf
```

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

```
File --> Load from File --> select /home/training/Desktop/untreated1.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 seven merged BAM files. E.g.:

```
samtools index ~/Desktop/untreated1.bam
```

Then try again to load the BAM file for the first sample into IGV.

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.

5 Ungapped transcriptome alignment

We shall now align a subset of our reads to the transcriptome rather than the genome. Open a terminal and `cd` to the `/nfs/training/ref` directory. The transcriptome FASTA and Bowtie 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?

5.1 Bowtie alignment

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

```
bowtie -a --best --strata -S -m 100 -X 400 --chunkmbs 256 --fullref -p 4 Dmel.BDGP5-transcripts \
-1 SRR031714_1.fastq -2 SRR031714_2.fastq | samtools view -F 0xC -bS - | \
samtools sort -n - ~/Desktop/untreated3-transcriptome
```

While this command runs, take a look at the Bowtie 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 ~/Desktop/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.

References

[1] Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R. 2011. Conservation of an rna regulatory map between drosophila and mammals. *Genome Res*, 21(2):193–202.