C S 489/509 Project 1
Comparing ovary and testis transcriptome of the fruit fly

Spring 2022

1 Introduction

The objective of this project is to assemble the transcriptomes of ovary and testis tissues of the fruit fly from RNA-sequencing data. Although ovary and testis cells have the same fruit fly genome, they perform sharply contrasting function. Ovary tissues are found only in female and testis tissues are only found in male fruit flies, respectively. Despite sharing the same genome, tissue differences arise from the distinct transcriptome as the consequence of a gene regulation program that runs differently in ovary versus testis.

This project can help answer the following biology questions. How does the transcriptome differs between ovary and testis in fruit fly? What genes are most characteristic of ovary and testis? Are there genes that are commonly expressed in both tissues?

2 An observational study of gene expression in fruit fly

Smibert et al. (2012) sequenced the transcriptome of various tissues of *Drosophila melanogaster* (Dmel), including ovary and testis. We pick these two tissue types to form a sharp contrast as their functions are highly distinct.

2.1 RNA-sequencing data

Testis RNA-seq data. Strand-specific polyA-enriched RNA-seq data were collected on *Drosophila* testis tissue from mated adult males at eclosion + 4 day. The dataset contain two replicates and can be downloaded at

`https://www.encodeproject.org/experiments/ENCSR254JFC/`

Replicate 1 has two paired-end read files in the fastq format, of size 500MB$\times$2. Replicate 2 has two fastq files of size 400MB$\times$2.

Ovary RNA-seq data. Strand-specific polyA-enriched RNA-seq data were collected on *Drosophila* ovary tissue from virgin adult females at eclosion + 4 day. The dataset also contains two replicates at

`https://www.encodeproject.org/experiments/ENCSR272DXE/`

where replicate 1 contains two fastq files of size 730MB$\times$2 and replicate 2 has two fastq files of size 510MB$\times$2.

2.2 Fruit fly reference genome and annotation

The current reference genome and annotation of *Drosophila melanogaster* (Dmel) is Release 6.44 announced in 2022 via FlyBase (FB2022_01). This release contains 32,074 gene records, where 17,874 genes are located to the genome and 14,200 are not located to the genome.

You can download the genome sequence file from the following URL:

`https://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.44_FB2022_01/fasta/dmel-all-chromosome-r6.44.fasta.gz`

You can download the genome annotation file in GFF3 format at

`https://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.44_FB2022_01/gff/dmel-all-r6.44.gff.gz`

or in GTF format at

`https://ftp.flybase.org/genomes/Drosophila_melanogaster/dmel_r6.44_FB2022_01/gtf/dmel-all-r6.44.gtf.gz`

StringTie seems to work with only the GTF file, despite its claim of supporting both GFF3 and GTF.

3 Analytical tasks

3.1 Transcriptome assembly using Dmel reference genome

(25%) Task 1. Transcriptome assembly. Map reads to the reference genome using HISAT2. The sequence alignment SAM file should be converted to BAM file format using

`samtools view`

to save space.

Try the two possible parameters `FR` and `RF` for strand specificity `–rna-strandness <string>`. Pick the one that results in the largest number of paired alignment.

(25%) Task 2. Transcript quantification. Quantify read counts per gene using StringTie. Report the numbers of both known genes and de novo genes. Use the option

`-G <ref_ann.gff>`

to utilize the genome annotation to flag out known genes or transcripts.

Report abundance for all genes in the genome in each sample.

Report abundance for all transcripts in the genome in each sample.

(25%) Task 3. Genes and transcripts of high fold-change. The log fold change of gene g is defined as follows:

$$r_g = \log_2 \frac{1 + \bar{g}_{\text{ovary}}}{1 + \bar{g}_{\text{testis}}}$$

where $\bar{g}_{\text{ovary}}$ is the average gene expression in the two replicates of ovary, and $\bar{g}_{\text{testis}}$ is the average for testis.

Find out the top five genes and transcripts with the greatest log fold change r_g between ovary and testis and top five with the lowest r_g.

Visualize the expression for these top genes and transcripts from the raw count data.

4 Extra credits (50%)

Smibert et al. (2012) used this dataset to study alternative polyadenylation in *Drosophila*. For those differentially expressed genes identified from your analysis, can you use some other tools such as Integrated Genomics Viewer (IGV) (`https://software.broadinstitute.org/software/igv/`) to visualize the read coverage of a gene and determine if there is any evidence for alternative polyadenylation for that gene between the testis and the ovary?

5 Project team

This project can be implemented as a team of up to two students of different academic backgrounds. Two students from identical majors are generally not allowed to be in the same group. Weekly progress must be reported in class until the project is due.

6 Project submission

(25%) Your project report will be a short paper organized as follows:

1. Title (≤15 words)
2. Abstract (≤250 words)
3. Introduction
4. Results. Answer questions raised in the tasks.
5. Data source.
6. Methods. Explain the experimental design. Please describe in your report the commands to rerun your code to reproduce the results.

   Document software version numbers and parameters that you used in the analysis.
7. Discussion. Describe the significance of the results, any challenges you have encountered, and new ideas to extend the work.
8. Conclusions. Summarize the meaning of your work and its significance.
9. Individual contribution. Clearly describe the tasks done by each member in the group. This section is not required for single-student groups.
10. References.

The project report and source code files must be submitted online by the due date.

References

Smibert, P., Miura, P., Westholm, J. O., Shenker, S., May, G., Duff, M. O., Zhang, D., Eads, B. D., Carlson, J., Brown, J. B., Eisman, R. C., Andrews, J., Kaufman, T., Cherbas, P., Celniker, S. E., Graveley, B. R., and Lai, E. C. (2012). Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep*, 1(3):277–289.