Creating an RNA-seq Differential Expression Analysis Tutorial Using Galaxy

AR Tilden – Colby College

Purpose:

- Create a **custom dataset** from publicly available RNA-seq data,
- **Small** enough to be run, start-to-finish, in the course of one afternoon lab period or a one-day workshop,
- That can be run on **public Galaxy** instance (this is variable, and Tophat may be slow on occasions).

Background:

If you're not Galaxy-familiar, run through the Mini-Mouse tutorial first to work through the basic RNA-seq analysis mechanics. You'll need a Galaxy account, and you'll need to download IGV (Integrated Genome Viewer).

Select Data to Analyze:

1. **Go to European Nucleotide Archive – Sequence Read Archive (ENA SRA): http://www.ebi.ac.uk/ena**
2. **Enter search terms RNA-seq paired end mouse (or whatever organism...).**

   Caveat: This tutorial assumes you'll use a species for which a genome annotation is available in iGenomes: https://support.illumina.com/sequencing/sequencing_software/igenome.html. If you're interested in a species not available in iGenomes, you'd need to dig deeper to find gene annotations – recall that this file allows you to tack gene symbols (names) to your data.
3. **Browse ENA SRA results for data with good experimental underpinning (in the case of Mini-Mouse, I selected PRJNA68307 on Eya2 knockout mouse: retina + control/Eya2 KO + midnight datasets + forward and reverse reads). Write down the accession numbers for your data; you'll likely want to go back to the files, or send your students there.**
4. **Upload to Galaxy via FASTQ files (Galaxy).**

Upload a Gene Annotations File:

1. **Reference Genome: Go to Illumina's iGenomes: https://support.illumina.com/sequencing/sequencing_software/igenome.html**
2. **I chose GRCm30 from Ensembl, downloaded it (took a while), unzipped tarball file, took just the iGenomes annotation file (a small file) and deleted the rest (a massive file!). It can probably be found in lots of other places, but I wasn't patient enough to look that hard...**
3. **Upload the annotations file into Galaxy.**

Data Formatting and RNA-seq Analysis:

1. Uploaded data are in a compressed format and need to be converted. Go to NGS: QC and manipulation → FASTQ Groomer → select your data and Input FASTQ quality score type: Sanger & Illumina 1.8+ (works for most newer datasets).
2. Go to NGS: QC and manipulation → FASTQC just to check in on quality scores; no need to trim data.
3. Go to TopHat, map reads to mm10 mouse genome or appropriate species. If data for fragment length are not available in the ENA SRA entry for your data, use defaults.
4. Check on TopHat alignment summary, quantity of accepted hit data.
5. Run Cufflinks with TopHat accepted hits, for both data pairs, using iGenomes annotations.
6. Check gene or transcript expression files just to make sure you've got data.
7. Run Cuffmerge on all data, using iGenomes annotations.
8. Run Cuffdiff on the 2 datasets (select control as first set, experimental as second).
9. Click on Cuffdiff: gene differential expression (click on name, not eye).
10. Click on the download icon. Open this txt file in TextWrangler (Mac), select all, and paste into Excel; results should be column-delineated (if not, go to Data → text to columns → ...).
11. Create a fold-change column after fpkm values (H - control and I - experimental): =h#/i#. This is optional, but it may be useful to eyeball fold-change unless you have a facility for intuiting $\log_2$ values.
12. Click to highlight second row, click Window → Freeze panes.
13. From here, there are a number of ways of looking at and sorting the data. GOAL: find a contiguous set of genes on a single chromosome where a handful (10-20) of genes are significantly up- and/or down-regulated. Sort on Column O – significant, Z-to-A, to isolate Yes values.
14. Select just Yes values and sort on locus. IF YOU'RE LUCKY, you may find a cluster of genes on one region of a chromosome – even 2 genes is good – that are up-/down-regulated. Keep in mind that there can be good data in the significant: No region as well. For example, if either the control or experimental is zero and the other column contains a high value, the result will not register as significant. So if you come up short, you could go back and sort by locus on all data and scroll through the chromosomes to find good regions. Ultimate goal: a group of 10-20 genes with some up/some down-regulation, some no difference; some with lots of expression, some with little.
15. Write down the coordinates that span this selected set of genes, and add a flanking region of several thousand up- and downstream of your coordinates.
16. It is useful to do a visual check-in with IGV at this point in the selected region to get an idea of what the data look like for the gene set. Here, you can get a nice visual of number of exons, mapped reads, and various other parameters as well – a variety is good in that regard.

17. Back in Galaxy: go to NGS: SAMtools → filter SAM or BAM, output SAM or BAM → select the two TopHat accepted hits files, and in Select regions, enter your chromosome range as, for example, chr1:12330000-12440000.
18. In NGS: Picard, select SamToFastq, and select filtered datasets. Counterintuitively, don't select Yes for "Do you want to output a fastq file per read group..." You will still get a separate forward (READ1) and reverse (READ2) dataset in the results. THE RESULTS will be your mini-dataset of reads. Assign appropriate names: xxxxforward/xxxxreverse.fastq.

Create a Gene Annotations file for just your chromosome of interest:

1. In the Galaxy tool panel, select Filter and Sort → Filter data on any column using simple expressions → Filter on (iGenomes annotation), with following condition ex. c1=='chr5.'

2. In the results, click on the pencil icon to change the name to something relevant; ex. iGenomes chr5 Annotations mm10. You could also create a more restricted chromosome range for just your genes of interest, but this is unnecessary and pedagogically too restrictive.

3. Now run the dataset through the Mini-Mouse tutorial to make sure everything works.

Sharing Data:

1. Go to View All Histories (right corner of Galaxy) → Create new (right corner).

2. Take your mini-dataset of reads and drag them into the new history; also drag your new annotation files over. Name the History.

3. Click on Analyze data to make this history data active. In this form, the dataset is ready to share:

4. Click the History options wheel in the History panel and select Share or Publish, and either add emails, share using a URL, or publish (the latter is the simplest, and it can be undone after everyone has loaded the data if you don't want your data accessible publicly). Alternatively, these are fairly small files; they can be downloaded or shared in any venue for sharing a standard-sized document. If downloaded → uploaded, the fastq files may need to be run through FastQ Groomer as described previously. You can check this by seeing if the files appear as options in Trimmomatic.