A Guide to Performing Differential Expression Analysis Using Galaxy Tools

Identifying and Visualizing Differential Gene Expression in Mouse Limbs Between Developmental Growth Stages

Karina Calderon

BIOINFORMATICS FALL 2019

A Guide to Performing Differential Expression Analysis Using Galaxy Tools:

Identifying and Visualizing Differential Gene Expression in Mouse Limbs Between Developmental Growth Stages

Link to Accompanying Video:

https://studentuml-my.sharepoint.com/:v:/g/personal/karina_calderon_student_uml_edu/EUANLGU8zmFLoHExkz8wYnkBJX0RWTWwWLJnhkG77inHKw?e=ldUriI

Objectives:

To identify differentially expressed genes between embryonic growth stages in mouse forelimbs, specifically between mouse forelimb tissues at E9.5 and E12.5 developmental stages using Galaxy Tools. The differential analysis is also visualized using volcano plots and heat maps.

Dataset:

RNA seq raw fastq files were found on NCBI/SRA. The RNA samples were extracted from mouse embryonic forelimb tissues at E9.5 and E12.5 developmental stages (4). **Single-end RNA-seq** libraries were prepared using ~100 µg of each of the RNA samples extracted from mouse forelimb tissues, then library quality control was performed, and libraries were sequenced with Illumina HiSeq 1500 (4). There were three biological replicates for each developmental stage (E9.5 and E12.5). Differential expression analysis can be performed using these raw fastq files.

https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP219910

Background:

Every cell within a multicellular organism contains the same genetic information (DNA), however, only a certain subset of genes is selectively “differentially expressed” in each cell depending on cell type, and on certain conditions (1). During development, different cells/tissues express different transcripts, and the specific combination of genes that are expressed or repressed dictates cellular morphology/function and can help reveal developmental, pathological, and other molecular mechanisms (1,2).

A GUIDE TO PERFORMING DIFFERENTIAL EXPRESSION ANALYSIS USING GALAXY TOOLS

Performing differential expression analysis can be a step in developing an increased understanding of specific gene functions. For example, we can ask “what genes are important for limb patterning and limb development between developmental growth stages and what are their specific functional roles?” Some first steps that can be taken to address this might be: sampling limb tissues from mice at various embryonic stages, extracting RNA, and conducting RNA seq. Then, using the resulting RNA seq data to perform differential expression analysis and genes that were upregulated in the various developmental stages may be identified. Those genes that were found to be upregulated might be playing a specific role in the patterning of limbs across embryonic development.

Although differential analysis cannot completely answer the question posed, again, it is one step towards developing subsequent experiments that can potentially further address the question which is “what genes are important for limb patterning and limb development between developmental growth stages and what are their functions?”.

Ultimately, RNA seq and bioinformatics approaches can be coupled with other techniques such as immunostaining, mouse genotyping, ChIP-qPCR, luciferase assays and whole-mount in situ hybridization, among other techniques can be used to explore this, as well as a myriad of molecular and cellular mechanisms.

This tutorial will be a simple guide to performing basic differential expression analysis between two groups of samples using tools found on Galaxy. According to Galaxy itself: “Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biological research” (3). Overall, the main advantage to using Galaxy and the tools it provides is that no coding experience is necessary making it easier for the novice to perform bioinformatics related analysis. Another advantage to using Galaxy is that all work is saved online and no software downloads are required. One disadvantage is that at times, the website may crash, unpredictably becoming unavailable.

A GUIDE TO PERFORMING DIFFERENTIAL EXPRESSION ANALYSIS USING GALAXY TOOLS

Tutorial Outline:

- Start by obtaining fastq files containing RNAseq reads
- Perform quality control (trimming reads if necessary) after running FastQC and MultiQC
- Mapping reads to reference genome with HiSAT2
- Quality control on mapped reads*
    - MultiQC, SamTools Stats, QualiMap BAM QC, QualiMap RNA-seq QC, QualiMap Counts QC
- Use featurecounts tool to count number of reads for each gene
- Create a matrix to aggregate count information for multiple samples with Column Join
- Prepare gene annotations
- QC on gene counts*
- Use Limma to perform differential expression analysis and plot results for visualization using volcano plots or heat maps
- Interpret results*

*Not fully covered in this general tutorial

How to find and upload RNA seq raw fastq files to Galaxy:

1. Go to NCBI: https://www.ncbi.nlm.nih.gov/
2. Open drop-down menu (by the search bar)
3. Select GEOdatasets
4. In search bar, type: mouse forelimb RNA seq /or ("replace with term of interest" RNA seq)
5. In the filters to the left of the page, select "samples"
6. Look through search results
7. Find sample of interest, and see if it is publicly available on SRA (or elsewhere)
8. Click SRA run selector (below sample name)
9. On SRA run selector, click on the Run ID (eg: SRR10046071)
10. Click on Data Access
11. Look where it says, "Original Format", find link for fastq file and copy link:

List of links used in this tutorial:
https://sra-pub-src-2.s3.amazonaws.com/SRR10046071/RNA-seq_E125_3.fq.gz.1
https://sra-pub-src-2.s3.amazonaws.com/SRR10046070/RNA-seq_E125_2.fq.gz.1
https://sra-pub-src-2.s3.amazonaws.com/SRR10046069/RNA-seq_E125_1.fq.gz.1
https://sra-pub-src-2.s3.amazonaws.com/SRR10046062/RNA-seq_E095_3.fq.gz.1
https://sra-pub-src-2.s3.amazonaws.com/SRR10046061/RNA-seq_E095_2.fq.gz.1
https://sra-pub-src-2.s3.amazonaws.com/SRR10046060/RNA-seq_E095_1.fq.gz.1

12. Open new tab with European Galaxy: https://usegalaxy.eu/
13. On the menu to the left click on the upload symbol (next to the star)
14. Select collection then select Paste/Fetch Data
15. Paste the links that were found to be associated with the samples of interest. Note: At least two different samples are required to perform differential expression analysis
16. Once all links are added to the list, click "start" to begin uploading the files to the galaxy history
17. Click build, and name the list by typing in "Mouse Forelimb RNA seq (raw fastq)"
18. Wait until uploading completes
19. The list/collection should appear under history

Run FastQC Reports:

FastQC Reports can be generated to assess fastq, BAM, or SAM files on the basis of the following elements:

It can also be used to identify presence of known adapter sequences in a file.

FastQC outputs multiple easy to read and understand graphs.

1. In Galaxy, on the left-hand tool search bar, search for and select FastQC
2. Under "Short read data from your current history" click the folder icon, this allows for selection of a collection, click on the collection name "Mouse Forelimb RNA seq (raw fastq)"
3. Scroll down and click on execute
4. Wait for reports to generate

Summarize FastQC Reports using MultiQC:

MultiQC is useful because it generates a single easy to view webpage that neatly summarizes the analyses of multiple samples.

1. In Galaxy, on the left-hand tool search bar, search for and select MultiQC

2. Under “which tool was used to generate logs” type or select “fastqc”
3. Under “select type of fastqc output”, select “raw data”
4. Then, under “FastQC output”, click the folder icon, select the raw data collection, it will say for example “10: FastQC on collection 8: Raw data”
5. Scroll down and click on “execute”
6. Wait for reports to generate

View MultiQC reports:

1. Once reports generate, under history, click on the eye icon next to the webpage MultiQC output
2. Assess report, determine if reads need to be trimmed, or adapters need to be removed

Map to reference genome using HISAT2 or Bowtie2:

Overall alignment rate was greater when using HISAT2, so we will be using HISAT2.

1. In Galaxy, on the left-hand tool search bar, search for and select HISAT2
2. Under “source for the reference genome”, select “Use a built-in genome”
3. Under “select a reference genome”, type in “mm10”, and select Mouse (Mus Musculus): m10 Full”
4. Under “Is this a single or paired library”, select “single-end”
5. Under FASTA/Q file, click the folder icon, and select the collection we created “Mouse Forelimbs Fastq (raw data)
6. Click summary options, and hit “yes” for the option to “output alignment summary in a more machine-friendly style” and for “print alignment summary to a file”
7. Then hit “execute”

Post mapping to reference genome, several quality control steps can be taken, however these will not be described in this tutorial.

Using featurecounts tool to count the number of reads for each feature (or chromosomal location):

Here we will generate files that have entrez geneIDs and corresponding numeric values indicating the number of reads found to map to the location of the respective geneID (for each sample). If there is no built-in genome for the organism which you are using data from, then the reference genome must be uploaded separately.

1. In Galaxy, on the left-hand tool search bar, search for and select featureCounts
2. Click the folder icon, and select the HISAT2 aligned reads BAM output
3. Under “gene annotation file”, select featurecounts built-in
4. Under “select built-in genome”, select “mm10”
5. Hit execute

Use Column Join on collections to aggregate the featurecounts Counts output into a matrix:

1. In Galaxy, on the left-hand tool search bar, search for and select Column Join on collections
2. Click the folder icon, and select the featurecounts Counts output
3. Under “identifier column”, type “1”
4. Under “Number of header lines in each input file”, type “1”
5. Under “Add column name to header”, click “no”
6. Hit execute

Edit matrix (“or table”), to simplify identifiers in header:

The purpose of this step is to simplify the identifiers in the header, as well as to avoid future errors. The original output file (if following steps in this tutorial) will contain the fastq links (that we uploaded) as the identifiers.

1. Download Notepad++ : https://notepad-plus-plus.org/downloads/
2. Download the matrix output from column join to your computer
3. Open the downloaded file with notepad ++
4. Relabel first column (first line) to say “Geneid”
5. Label columns 2-6 respectively as E095_1, E095_2, E095_3, E125_1, E125_2, E125_3 (no commas)
6. Delete the last line of the table. This was necessary because the last line also contained the fastq links that we are using for this analysis, which are non-numeric values. The table should only contain a header (the first row), the GeneIDs in the first column, then numeric values corresponding to the number of reads found at the locations of each gene. The presence of additional non-numeric values will result in an error in future steps.
7. Save the edited file, renaming it: Columnjoinedited.tabular
8. Upload edited file to Galaxy

Get gene annotations using annotateMyIDs tool:

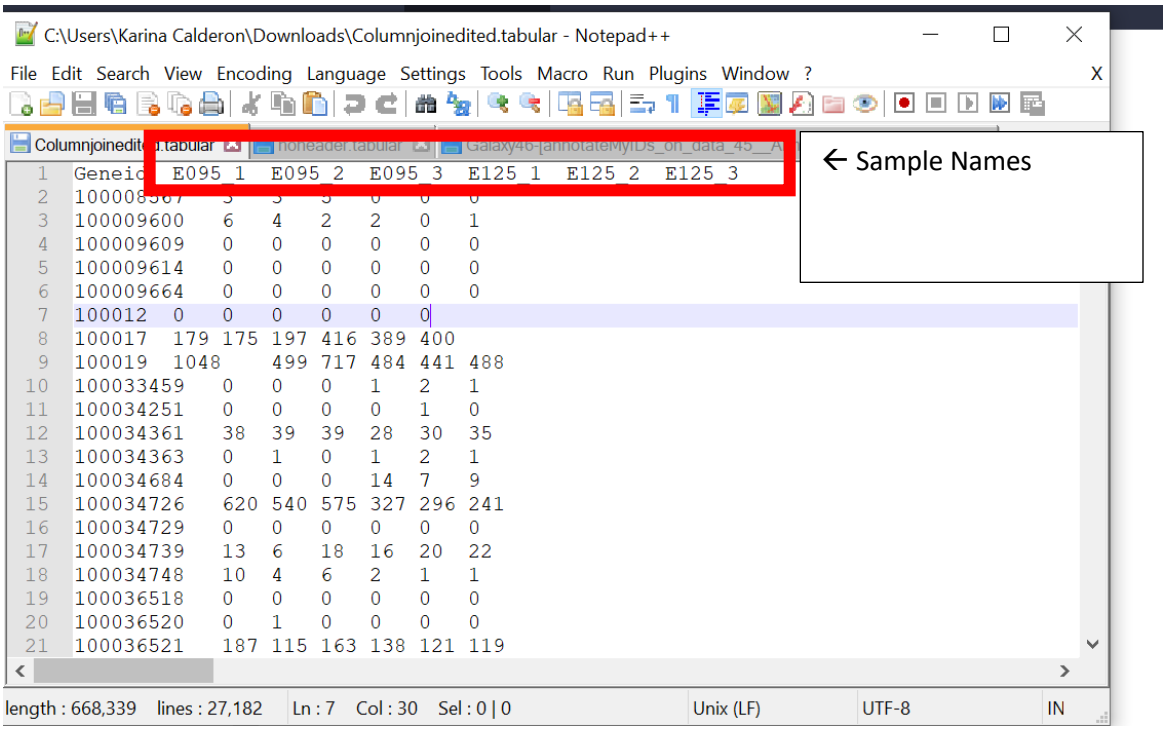
This step will add the gene names, and symbols that correspond to the entrezIDs (which are only numeric IDs).

1. In Galaxy, on the left-hand tool search bar, search for and select annotateMyIDs
2. Click on the page icon, then select the Columnjoinedited.tabular file
3. Under “file has header?” click “Yes”
4. Under “organism”, select “mouse”
5. Under “ID type”, select “Entrez”
6. Under “output columns”, mark ENTREZID, SYMBOL, and GENENAME
7. Hit execute
8. Once the job is complete, click the eye icon to view the annotatemyIDs output, under the file name, check the number of lines, making sure that the number of lines for the annotation output file, is the same as the input Columnjoinedited.tabular file.

annotation file and matrix (column join) files should have the same number of lines!

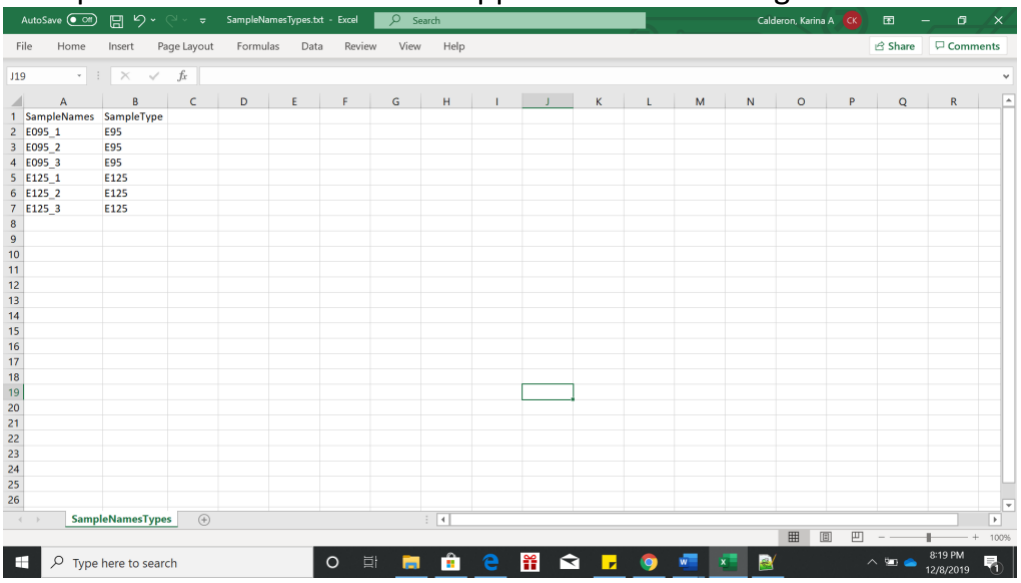
Create a table with sample names and sample types (factor information file):

1. Open a new spreadsheet in excel, and the Columnjoinedited.tabular file (in notepad ++)
2. In cell A1 on excel, type “SampleNames”, in cell A2 type “SampleType” (excluding the commas)
3. Under “SampleNames” type in the sample names as according to the the header of the Columnjoinedited.tabular file (must be in the same order in both files (ie: E095_1, E095_2, E095_3, E125_1, E125_2, E125_3)



	E095_1	E095_2	E095_3	E125_1	E125_2	E125_3
100009600	6	4	2	2	0	1
100009609	0	0	0	0	0	0
100009614	0	0	0	0	0	0
100009664	0	0	0	0	0	0
100012	0	0	0	0	0	0
100017	179	175	197	416	389	400
100019	1048		499	717	484	441	488
100033459	0	0	0	1	2	1
100034251	0	0	0	0	1	0
100034361	38	39	39	28	30	35
100034363	0	1	0	1	2	1
100034684	0	0	0	14	7	9
100034726	620	540	575	327	296	241
100034729	0	0	0	0	0	0
100034739	13	6	18	16	20	22
100034748	10	4	6	2	1	1
100036518	0	0	0	0	0	0
100036520	0	1	0	0	0	0
100036521	187	115	163	138	121	119

4. Under “SampleType”, type in E95 for the first three rows, and E125 for the last three rows
5. The spreadsheet in excel should appear as the following:



SampleNames	SampleType
E095_1	E95
E095_2	E95
E095_3	E95
E125_1	E125
E125_2	E125
E125_3	E125

6. Then name the file: SampleNamesTypes
7. Save the file as a tab delimited text file
8. Upload it to Galaxy

At this point we have three main files:

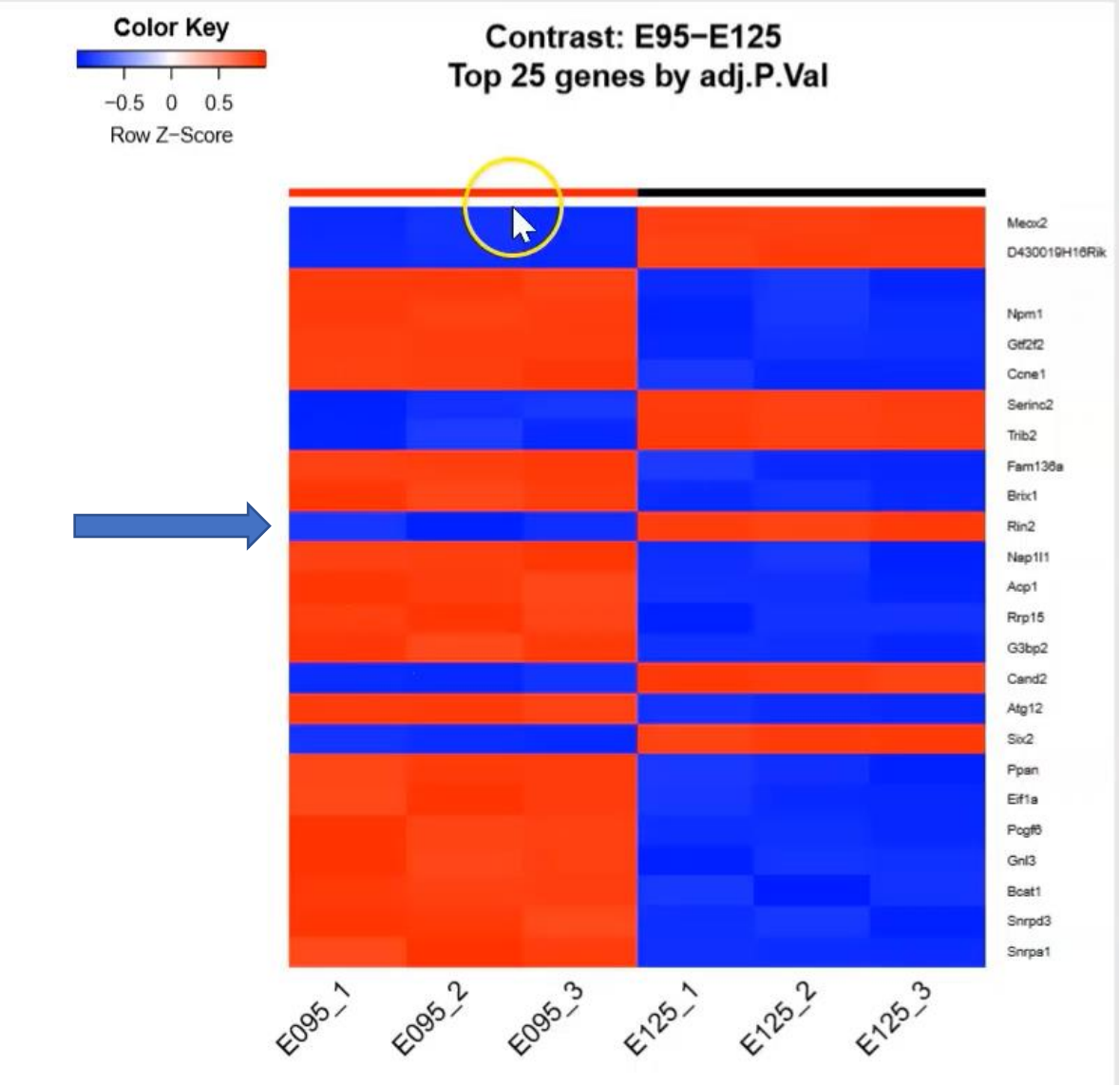
- Gene Annotation File
- Matrix (Column Join) file
- File with sample names/types (factor information file)

Now, we can run Limma

Run Limma to perform differential expression analysis via limma-voom :

1. In Galaxy, on the left-hand tool search bar, search for and select Limma
2. Under differential expression method, select limma-voom
3. "Apply voom with sample quality weights?" click "No"
4. Under "Count files or matrix?", select "Single Count Matrix"
5. Click the page icon and select our "columnjoinedited.tabular" file
6. Under "input factor information from file?", click "yes"
7. Under factor file, click the page icon, then select the "samplenamestypes.txt" file that we created
8. Under "use gene annotations" select yes, click the page icon, the select the "annotatemyIDs" file
9. Under "input contrast information from file?" select "No"
10. Under "contrast of interest" type in the contrast of interest, can be for example "E95-E125" or "E125-E95"
11. Then click on "filter low counts"
12. Under "filter lowly expressed genes" select "yes"
13. Under "filter on CPM or Count Values?", select "CPM"
14. Under "minimum CPM" type in (for example) "6"
15. Under "minimum samples", type "2"
16. Click on "output options"
17. Under "additional plots", select all
18. Click on "advanced options" and under "minimum log2 fold change", type in (for example) "2"
19. Scroll down to "number of genes to highlight in Volcano plot, Heatmap and Stripcharts", type in (for example) "25"
20. Hit execute
21. Wait for job to finish, then view and interpret the output files

Eg: Output (HEATMAP)



Based on this output (given the filtering we performed and the parameters we set), we can say for example that the gene “Rin2” was expressed less in the E9.5 forelimb samples relative to the E12.5 forelimb samples. The blue suggests lower expression

while the red suggests higher expression. One can then go on to review the significance of the differentially expressed genes, and possibly perform follow up experiments. For example, we can read about the function (if known) of the "Rin2" gene. We can read on NCBI, that mutations in this gene can result in disorder of elastic/connective tissue and phenotypes such as sagging skin, macrocephaly, alopecia, cutis laxa, and scoliosis (6).

*Note, the parameters set in this tutorial are not "ideal" parameters. Every set of data will require a different set of parameters, and filtering steps. One can learn more about the individual tools used in this tutorial and their options by reviewing other resources.

References:

(1) Francesca Finotello, Barbara Di Camillo, Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis, Briefings in Functional Genomics, Volume 14, Issue 2, March 2015, Pages 130–142, https://doi.org/10.1093/bfgp/elu035

(2) Ralston, A. & Shaw, K. (2008) Gene expression regulates cell differentiation. Nature Education 1(1):127

(3) https://usegalaxy.eu/

(4) https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP219910

(5) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4049934

(6) https://ghr.nlm.nih.gov/gene/RIN2#conditions

(7) https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/differential-gene

(8) Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015(11):951–969. Published 2015 Apr 13. doi:10.1101/pdb.top084970