**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**Lab 6 Microarray and RNA-seq Analyses**

**Dr. Arun Sethuraman**

**San Diego State University**

The goal of this week’s lab is to analyze and estimate differential gene expression, as studied using two “high-throughput” techniques, namely microarrays and RNA-seq. Microarrays (either using mRNA or cDNA target/probe mechanisms) and RNA-seq (sequencing entire transcriptomes using high-throughout next generation sequencing) have revolutionized our understanding of gene expression and broadly, of functional genomics (i.e. understanding the functions of genes, and genomes). Both techniques utilize a “case-control” type of experimentation to either directly estimate (a) the number of transcripts that are expressed/produced under each experimental condition, or (b) the relative degree of hybridized transcripts that are a proxy for differential gene expression. Both methods have proven to be extremely useful in a variety of fields, including medicine (e.g. differential expression in cancer versus non-cancerous tissue), understanding physiological response to environmental variation (e.g. drought tolerance in plants), tracing differential gene expression in time-dependent life-history stages (e.g. differentiation in embryonic versus zygotic tissue). Some of the differences in the methods are highlighted in Table 1.



Table 1 Bauer et al. 2014 <https://doi.org/10.1186/1471-2105-15-S11-S3>

**Part 1: Microarray Data Analyses** (adapted from <http://barc.wi.mit.edu/education/bioinfo2007/arrays/index.html>)

1. As described in [Su et al., 2002](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=11904358" \t "_blank), human tissue samples were hybridized on Affymetrix (one-color) arrays and chips were scanned. For each tissue, at least two independent samples were hybridized to separate chips.
2. Scanned images were quantified (including measurement of background) using standard software.
3. Data was normalized, and log-transformed for this analysis today.
4. Data was also filtered to only contain data with non-questionable expression levels.
5. To obtain this transformed dataset go to Canvas and under Lab 6 download “exprset\_present.txt” to your JetStream page.

Now in R:

exprSet = read.table("exprset\_present.txt")

# Check how the chips are named

colnames(exprSet)

1. Use the t test for one gene to determine if the data on fetal and adult expression are different in the brain and/or liver. We'll use the command "t.test" to perform Welch's test (not assuming an equal variance between samples) on the data for one gene (probset). We're safer by not assuming that each genes exhibits the same variance in each tissue, but we do lose some statistical power. The two-sample t-test takes at least three arguments: the first dataset, the second dataset, and an indication of the number of tails. For 2 tails, we'll state "two.sided", since one tissue can have an expression that is lower or higher than the other.
2. The command to compare adult brain to fetal brain for the first gene is

dataset.1 = exprSet[1, c("brain.1", "brain.2")]

dataset.2 = exprSet[1, c("fetal.brain.1", "fetal.brain.2")]

# or we can indicate columns by their numbers

dataset.1 = exprSet[1, c(1,2)]

dataset.2 = exprSet[1, c(3,4)]

t.test.gene.1 = t.test(dataset.1, dataset.2, "two.sided")

# Let's see what these data are

dataset.1

dataset.2

# Print just the p-value from the t-test

t.test.gene.1$p.value

3. Use the t test to test for a difference in means with all genes. Note the second argument is "1", showing that we want to apply the t-test command across rows (genes/probesets)

brain.p.value.all.genes = apply(exprSet, 1, function(x) { t.test(x[1:2], x[3:4]) $p.value } )

liver.p.value.all.genes = apply(exprSet, 1, function(x) { t.test(x[5:6], x[7:8]) $p.value } )

# Check the first few brain ones to make sure the first one agrees with our single-gene command

brain.p.value.all.genes[1:5]

4. Correct t-test p-values for multiple hypothesis testing by using three methods - Bonferroni, Sequential Bonferroni, and FDR. Note that the actual p-values representing confidence for differential expression are raw values. If they were to be corrected for multiple hypothesis testing (since were doing lots of t-tests), they'd be much higher.

#Bonferroni adjusted p-values:

brain.bonferroni.pvals=p.adjust(brain.p.value.all.genes,method=”bonferroni")

liver.bonferroni.pvals=p.adjust(liver.p.value.all.genes,method=”bonferroni”)

#Sequential Bonferroni (also called the Holm-Bonferroni method) adjusted p-values:

brain.seqbonferroni.pvals=p.adjust(brain.p.value.all.genes,method=”holm”)

liver.seqbonferroni.pvals=p.adjust(liver.p.value.all.genes,method=”holm”)

#FDR

brain.fdr.pvals = p.adjust(brain.p.value.all.genes, method=”fdr”)

liver.fdr.pvals = p.adjust(liver.p.value.all.genes, method=”fdr”)

5. Sort all of our unadjusted, Bonferroni, Sequential Bonferroni, FDR\_corrected p-values to get lowest values (commands here shown only for FDR - use the same method for unadjusted, Bonferroni, Sequential Bonferroni).

brain.fdr.pvals.sorted = brain.fdr.pvals [order(brain.fdr.pvals)]

liver.fdr.pvals.sorted = liver.fdr.pvals[order(liver.fdr.pvals)]

# Look at the 10 lowest p-values

brain.fdr.pvals.sorted[1:10]

liver.fdr.pvals.sorted[1:10]

6. List all the gene IDs for those that meet your significance threshold (such as unadjusted p < 0.01) and are present in at least one sample. You’ll notice that all the corrected p-values will not meet any significance threshold for this dataset, which means that in an ideal scenario, we would do many more replicate probeset runs.

x<-brain.p.value.all.genes[brain.p.value.all.genes < 0.01]

y<-liver.p.value.all.genes[liver.p.value.all.genes < 0.01]

Sort these two lists, and determine the top 10 “most significantly” differentially expressed genes between their respective fetal and adult tissues.

Now use the intersect command to find out genes that are differentially expressed in brain and liver tissues between fetal and adult tissues:

intersect(names(x),names(y))

**Part 3 - RNA-seq Analyses**

In today’s lab, we will focus on performing downstream data analyses on an RNA-seq dataset, adapted from Myrto Kostadima’s RNA-seq analysis workshop (November 2016). This tutorial has been uploaded to your Canvas as well (apart from the files we will be analyzing). This can be found under Lab 6 tab as “RNA-seq Tutorial from EMBL”. We will make use of this opportunity to do some NGS analyses using the publicly available Galaxy Project server.

The goal of this hands-on session is to perform some basic tasks in the downstream

analysis of RNA-seq data. We will start from RNA-seq data aligned to the zebrafish

genome using Tophat. We will perform transcriptome reconstruction using Cufflinks

and we will compare the gene expression between two different conditions in order to

identify differentially expressed genes using Cuffdiff.

We will use a dataset derived from sequencing of mRNA from Danio rerio embryos in

two different developmental stages. Sequencing was performed on the Illumina platform

and generated 76bp paired-end sequence data using poly-(A)+ selected RNA. Due to

the time constraints of the practical we will only use a subset of the reads.

The data files are contained in the subdirectory in your Canvas page under Lab 5 and are the following:

2cells\_1.fastq and 2cells\_2.fastq: these files are based on RNA-seq data

of a 2-cell zebrafish embryo.

To obtain these, click on the link “Link to download FASTQ files from” and on the webpage scroll down to where it says “Click here to download the worksheet and files”.

Or follow this link provided to download the FASTQ files.

<ftp://ftp.ebi.ac.uk/pub/training/Train_online/RNA-seq_exercise/>

2. Now let’s navigate to [www.usegalaxy.org](http://www.usegalaxy.org).

3. Now let’s upload our .fastq files onto the Galaxy server. Note: to do this navigate to the left panel and click “Get Data” and choose “Upload File”.

4. Now we will run TopHat on it (to assemble our data set by using the Zebrafish genome as a guide). Note: to do this navigate to the left panel and select “search tools” and type in “TopHat”. FYI: You will notice that this tool has been deprecated, meaning that the tool is less favorable than other tools/methods available for RNA-seq analysis. Click on “TopHat” and select “Paired-end (as individual datasets)” under “Is this single-end or paired-end data”. For forward reads choose “2cells\_1.fastq” and for reverse choose “2cells\_2.fastq. Execute TopHat (this might take a while to run, especially since we’re all submitting data at the same time). So perhaps it would be wise for you to work in pairs for this exercise. Do the same with the 6 hours files.

4. After TopHat completes, look at the output files that are produced. Visualize the obtained “hits” in one of the many genome browsers. To do this click on the different outputs and at the very bottom click on the bars symbol. Or click on the “eye” to see the numerical output. Use the chart button in the output, then select the Danio rerio reference genome. This should open a new browser window with your “hits”.

Exercise:

1) How many hits do you obtain against the reference genome? Will this change if you ran the alignment with TopHat using only one of the fastq files? Test this hypothesis.

2) Now search ENSEMBL for the CBY1 gene in the Danio rerio database. Obtain the coordinates for this gene. Search your genome browser track for these coordinates. Observe the Ensembl annotation, and RefSeq annotation. Are the splice junctions annotated? Are there any reads that match the intronic regions (between splice junctions)? Why/why not?

5. Now let’s analyze isoform expression and transcriptome assembly. There are a number of tools that perform reconstruction of the transcriptome and for this workshop we are going to use Cufflinks. Cufflinks can do transcriptome assembly either ab initio or using a reference annotation. Cufflinks has a number of parameters in order to perform transcriptome assembly and quantification. To view them all, use the Cufflinks link inside your Galaxy window, and explore the different options.

We aim to reconstruct the transcriptome for both samples by using the Ensembl annotation

both strictly and as a guide. In the first case Cufflinks will only report isoforms

that are included in the annotation, while in the latter case it will report novel isoforms

as well.

A reminder from the presentation this morning that FPKM stands for Fragments Per

Kilobase of exon per Million fragments mapped.

The annotation from Ensembl of Danio rerio is stored under the folder annotation is

a file called Danio\_rerio.Zv9.66.gtf.gz (in Canvas – go ahead and download this and upload it into Galaxy after you unzip it).

Now execute Cufflinks on the accepted hits bam files for both the conditions – 2 cells, and after 6 hours, using the gtf file as a reference annotation. To do this search for “Cufflinks” in the left panel under the search tools then choose your “accepted\_hits” for TopHat on your datasets under “SAM or BAM file of aligned RNA-Seq reads”. Then under “Use Reference Annotation” select “Use reference annotation” and choose “Danio\_rerio.Zv9.66.gtf

6. Take a look at the files that have been created.

Here’s a short description of these files:

a) Skipped transcripts – this GTF file contains Cufflinks isoforms/reads that were not assembled

b) Assembled transcripts - this GTF file contains Cufflinks assembled isoforms

c) Transcript expression - contains the estimated isoform-level expression values.

d) Gene expression - contains the estimated gene-level expression values.

The complete documentation can be found at: <http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/>

7. We will now repeat these analyses on a second set of files that have been provided for you, which were splice variants and mRNA obtained after 6 hours from the D. rerio samples.

8. Once this has completed, we will analyze differential expression. One of the stand-alone tools that perform differential expression analysis is Cuffdiff.

We use this tool to compare between two conditions; for example different conditions

could be control and disease, or wild-type and mutant, or various developmental stages.

In our case we want to identify genes that are differentially expressed between two

developmental stages; a 2 cell embryo and 6h post fertilization.

Open CuffDiff from the Galaxy pane, use both the assembled transcripts from the previous CuffLinks run (click on “Multiple datasets” in the “Transcripts” upload box, choose both assembled transcripts, as well as the reference annotation (GTF file provided)), and all the following options: use the BAM files created (accepted hits) for two conditions - 2 cell embryo versus a second condition - 6 hours, set “Use multi-read correct” to YES, Perform Bias correction using the reference genome fasta file (download this from Canvas, and upload it to your Galaxy), Set Advanced CuffDiff parameters->Library prep used for input reads should be set to fr-unstranded. Execute CuffDiff, and while it runs, read more about the output files here: <http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/#cuffdiff-options>

9. We are interested in the differential expression at the gene level. The results are

reported by Cuffdiff in the file “Gene Differential Expression testing”. To download this file scroll to the bottom and click on the floppy disk. Upload this file onto your server (biol480) and rename it to gene\_exp.diff

Look at the first few lines of the file using the following command:

head -n 20 cuffdiff/gene\_exp.diff

We would like to see which are the most significantly differentially expressed

genes. Therefore we will sort the above file according to the q value (corrected

p value for multiple testing). The result will be stored in a different file called

gene\_exp\_qval.sorted.diff.

sort -t$'\t' -g -k 13 cuffdiff/gene\_exp.diff >

cuffdiff/gene\_exp\_qval.sorted.diff

Look again at the first few lines of the sorted file by typing:

head -n 20 cuffdiff/gene\_exp\_qval.sorted.diff

Copy the Ensembl identifier of one of these genes. Now go back to the Ensembl genome browser <https://uswest.ensembl.org/index.html> and paste it in the search box. Look at the raw aligned data for the two datasets.

Copy just the Ensembl id column of the top 100 hits, and save this to a separate file, called “top100.genes”

Exercise:

Do you see any differences in gene coverage between the two conditions that would justify that this gene has been called as differentially expressed?

10. Functional Annotation of Differentially Expressed Genes:

After you have performed the differential expression analysis you are interested in

identifying if there is any functionality enrichment for your differentially expressed

genes.

Open a web browser and go to the following URL <https://david.ncifcrf.gov>

On the left side click on Functional Annotation. Then click on the Upload tab. Under

the section “Choose from File”, click “Choose File” upload the top100.genes file you just created.

11. Select “ENSEMBL\_GENE\_ID” from the drop-down menu of “Select Identifier”. Finally select “Gene List” and then press “Submit List”.

Click on “Gene Ontology” and then click on the “CHART” button of the “GOTERM\_BP\_ALL”

item.

Questions

Do these categories make sense given the samples we’re studying?

Browse around DAVID website and check what other information are available.