Homework 3: E­­xpression analysis

Full Names:

Group:

**Form:**

You have to supply both the answer (whatever it is – numbers, a table, plots or combinations thereof), as well as the R code you used to make the plots. Note that

i) If the R code gives different results than your results, you will get severe point reductions or even 0 points for the exercise.

ii) Some questions will request you to use R options/functions we have not covered explicitly in the course: this is part of the challenge.

iii) While this is a group work, we expect that everyone in the group will have understood the group solution: similar or harder question might show up in the individual homework (exam). So, if something is hard, it means you need to spend more time on it.

The assignment should be submitted as a pdf file – it is preferable to use the actual word document with the questions as a working document and then save as pdf. If your computer cannot save as PDF, consider downloading this small freeware program:

PDFCreator <http://sourceforge.net/projects/pdfcreator/>, which will show up as a “fake” printer that actually saves PDF files.

Plase make sure to include the group number and submission number in the file name

The total length of this homework cannot exceed 12 pages. The file should be uploaded on Absalon where you downloaded this file before the deadline.

In the solutions, you should describe briefly what you are doing, and motivating why you do it.

For statistical tests, you have to

1) Motivate the choice of test,

2) State exactly what the null hypothesis is (depends on test!)

3) Comment the outcome – do you reject the null hypothesis or not, and what does this mean for the actual question we wanted to answer (interpretation)?

A question marked \* means that it is more challenging, and likely requires skills from the whole group.

# Part 1 – Microarrays

1) Your colleague has made two microarray expression experiments: one for HIV-positive patients (5) and one for healthy controls (5). She is interested in two genes in particular (RXRA and IRX3).

Calculating the mean signal of these two genes across the patient groups (HIV and non-HIV) shows that IRX3 always has higher signals than RXR. Can we then conclude that IRX3 is higher expressed in all these samples? Explain why/why not?

2) The actual normalized data is in the file normalized\_data.txt. The five first columns are from HIV patients and the last five from healthy controls. Each row contains the normalized expression values from the probes(s) corresponding to one gene. Do a suitable statistical test for each row to find the differentially expressed genes (show the R code only – we will use the result in the next few questions)

3) How many false positives would you expect for this experiment if you use a threshold of 0.05? How many genes do you actually get with a p-value less than 0.05?

4) The function p.adjust(p-values) can be used to correct for multiple testing. How many genes do you get with a p-value <0.2 when you use the Bonferroni correction? How many do you get with a FDR(Use the BH method) than 0.2. How many of these genes(FDR<0.2) would you expect to be false positives?

5)\* She also want to see how big the changes between the conditions are. So calculate the log2 foldchange for each gene. You can use this formula:

foldchange=log2(mean(hiv))-log2(mean(control))

Hint: One way to do it use the apply function twice to first calculate two vectors with the mean values and then use the formula afterwards.

6) Report the fold changes for the genes with a FDR<0.2. Are there most up (Up in HIV) or down regulated genes in this subset? Comment on the size of the log2FCs

# Part 2 – Identification of isoform switching based on RNA-seq data

Use hg19 Feb 2009 for this assignment.

## Data analysis in Galaxy

Download the 6 sam files from absalon (WT\_1.sam, WT\_2.sam, WT\_3.sam, KD\_1.sam, KD\_2.sam, KD\_3.sam). A sam file is the non-binary version of bam-files (that we used in the class-exercise). These 6 files contain all the reads that mapped to a 500 kb region on chromosome 1.

This dataset uses sequencing data from <http://www.nature.com/nbt/journal/v31/n1/full/nbt.2450.html> and the files are a subset of the output of TopHat. In this experiment HOXA1 was knocked down in triplicates resulting in the 6 files (WT = Wild type = control) (KD = Knock Down).

1) Upload all 6 files to the BINF galaxy server (http://galaxy.bio.ku.dk/). Remember to use hg19 as reference genome. Import a GTF file containing all knownGenes (UCSC genes) in the region chr1 position 1,000,000 to 1,500,000. This should be used as a reference transcriptome in this exercise.

2) Use cufflinks to assemble the transcriptomes for the individual samples. Use the imported knownGenes as reference guide annotation and leave the rest at default parameters.

* Q: For each condition calculate the average number of lines in the 3 ”assembled transcripts”

3) Use Cuffmerge to obtain a single combined transcriptome from all the samples, and use the imported UCSC genes as Reference Annotation. Additional information about the run can be found by pressing the “I” button when the dataset is expanded.

* Q: How many lines does the combined GTF file have? How does that compare to the individual transcriptomes? What does this suggests?
* Q: Download the dataset by pressing on the download button to export the combined transcriptome. You should then load it to the UCSC genome browser (<http://genome.ucsc.edu/>) using My Data -> Custom Tracks tabs andake a screenshot (with default tracks) of the (whole) region with the data and include it in your answers. Use “squished” for knownGenes and your User Track.

4) Use “Cuffdiff” to make the differential expression analysis between the two conditions (WT and KD) (by highlighting the appropriate files). Turn on “multi-read correct” with default parameters.

* Q: While we are not using it here, why would be want to use bias correction (google it and include reference for answer).

5) After the CuffDiff run is done, remove all results except the result of the differential expression analysis on gene, transcript and splicing: ”gene differential expression testing”, ”transcript differential expression testing” and “Splicing differential expression testing.

* Q: Specify what CuffDiff have tested in the “Splicing differential expression testing”. Google or deduce it, include deduction and/or refrence.

6) To enable us to check your workflow you need to save html file showing your work flow. Please make sure to delete all non-used data from your history (for example you should remove all Cufflinks output not used (by pressing the “x”)).

The html file is generated as follows:

1. Make sure that all datasets in your history are expanded (by clicking them once).
2. In the history options menu (the small wheel at the top of your history) select ”Show structure”. This should open a site showing your galaxy workflow, including your settings.
3. When this site have loaded right click anywhere on the page, where the mouse icon does NOT change to a hand (no mouse-over effect - it should look like your standard curser). Choose ”save as”. Save the html file to your computer.
4. Check that the html file work by opening it in a browser – the result should look identical to your galaxy except none of the buttons can be used.

Upload the resulting HTML file AND the associated folder together with the answers and the 3 CuffDiff result files files to the homework (zip the html and the folder into one).

# Part 3 - Post analysis in R

Remember to report sufficient information and R code to enable us to replicate your results!

**For this assignment you need to use the 3 files that are provided along with the homework for the rest of the assignments (can be found in “data\_for\_part\_3” folder)!**

Read the supplied 3 CuffDiff result files (”gene differential expression testing”, ”transcript differential expression testing” and “Splicing differential expression testing”) into R as three data.frames (one for each). Note that the files provided are different from what you are expected to get if you solved part 2, so the results should not be compared.

1) The unique transcript id in the transcript data.frame is “test\_id” (which is also a column in the gene data.frame). Change the column name (“test\_id”) to ” transcript\_id” to enable us to differentiate between them.

2) Make two new data.frames that only contains genes/transcripts that are expressed in at least 1 condition. For each data.frame make this in one line of R code without using the semicolon (;). Use these data.frames in the rest of the assignment.

3) How many genes and how many transcripts were expressed? How many genes and how many transcripts were significantly differentially expressed between conditions?

4) Make two new data.frames (one for gene, one for transcripts) where the transcript data.frame only contains the transcript\_id, gene\_id, value\_1 and value\_2 columns and the gene data.frame only have the gene\_id, gene, value\_1, value\_2 columns.

Use the merge() function to combine these two data.frames, based on gene ids. Use the suffix parameter to make the resulting column names easily understandable. How many rows does this new data.frame contain? How many columns?

5\*) For all transcripts calculate the Isoform Fraction values (IF values) and the corresponding dIF values. Do any of these calculation results in NAs? Explain why you could get NAs and discuss whether this should be corrected (by for example setting it to 0 or 100).

6\*) What is the average (mean) and median dIF value? Compare the two values and discuss what it enables you to say about the distribution of dIF values.

7) Use R to subset the merged data.frame to only contain genes with potential isoform switching by identifying genes with dIF > +/- 0.25 (0.25 is an arbitrary (but large) value). Furthermore add the p\_value from the “Splicing differential expression testing” to the data.frame using the match() function.

8) For the switch in the gene with the lowest p\_value report

A) the transcript ids

B) the gene name (not the gene\_id), if the gene have multiple names just report the first.

C) The dIF values.

D) The pvalue. Include the R code to extract exactly this data.

9) Analyze the gene with a switch: What does the gene do (a few sentences with refrences). Take a look at the gene in the genome browser (upload and use the supplied GTF file instead of the one you made in part 2). Make sure to compare it to the knownGenes annotation you provided as guide. What is the difference between the transcripts involved in the isoform switches? Compare the alternatively spliced regions to the “Pfam in UCSC Gene” data track. Rearrange the tracks so the order is: 1) your track, 2) UCSC genes and 3) Pfam Domains. Report you findings including screenshots. Include an explanation of why it is (potentially) interesting to compare to the pfam track.