**Names:**

**Daniel Carvalho**

**Rafael Segura**

**Course:**

**Big Questions 2**

**All the answers are found in blue within this document.**

**THE assignment (fun!)**

**[shamelessly borrowed from STAT540 at UBC]**

In this assignment, the questions are structured to test your knowledge on some statistical concepts and R techniques, such as data visualization, data wrangling, etc. If you find yourself referring to the seminars and perhaps googling things a lot, you’re on the right track. Asking the TAs for help during the seminars is also a good option, too!

Also check out assignment\_tips.

**Evaluation**

Each student/group must submit their own work. It’s fine to talk to fellow students. If someone is really helpful, mention them in your homework to give them some credit. It’s crossing the line to copy and paste any code or wording.

**Submission**

Please refer to the assignment\_tips for submission instructions. It’s due on **April 28**. Late submission will simply annoy the instructor.

**About R session**

It might not matter but just in case: this assignment assumes that you have

* R version 3.2.4 (2016-03-10) – "Very Secure Dishes"
* Bioconductor 3.2
* limma 3.26.9
* edgeR 3.12.1

Note that limma, edgeR need to be download via bioconductor (eg biocLite("limma"); biocLite("edgeR")), the rest are installed by runnnig install.packages() (eg install.packages("gridExtra"), etc)

Also another tip: if you’re using RStudio, you can use tab to autocomplete any filepath, variable names, function names, etc.

For questions 1-5, you will be analyzing a publicly-available gene expression study of normal human bronchial epithelial (NHBE) cells, run on the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array. Please briefly read about the study: [Time course of NHBE cells exposed to whole cigarette smoke (full flavor)](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10718). We will use “probes” and “genes” interchangeably, and also “agent” and “treatment”.

**Question 1: Data inspection**

**1.1 Download and inspect the data**

The transcriptome data and the meta-data for the design details can be downloaded from Blackboard (NHBE\_transcriptome\_data.txt.gz and NHBE\_design.txt). The data has been preprocessed and it is on a log2 scale. One sample with NA values on all probes is removed from both the data and the meta-data, so overall it's a little different than the processed data provided via GEO.

* Please load and inspect the data and the meta-data. Fiddle with factor levels and do any other light cleaning you deem necessary.

Not cleaning was necessary after download and inspection of the data. Although we could change order of columns, not big deal.

* How many genes are in our data? How many samples?

22737 genes

* Check the breakdown for treatment and time. Do you think it’s a good experiment design?
  + Hint: table() and addmargins() will help.

We used the commands table and str() to see the main features of the table and we see that the experimental design is fine. We just noticed that not all the time points have the same number of replicates, which might represent a problem for some analysis. The cartoon below summarizes the experiment.

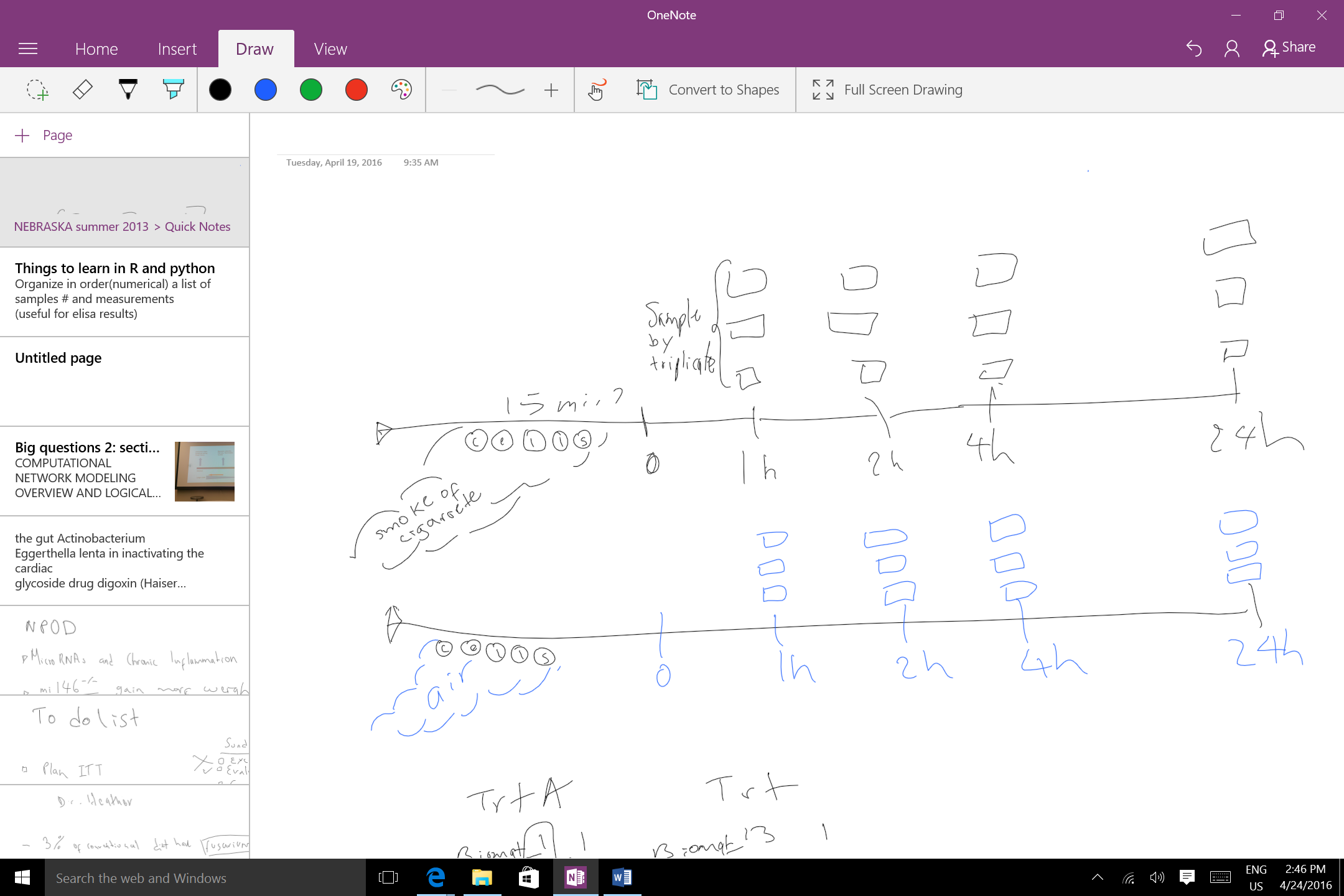


Figure 1. Experimental and Treatment Design.

**1.2 Basic data manipulation**

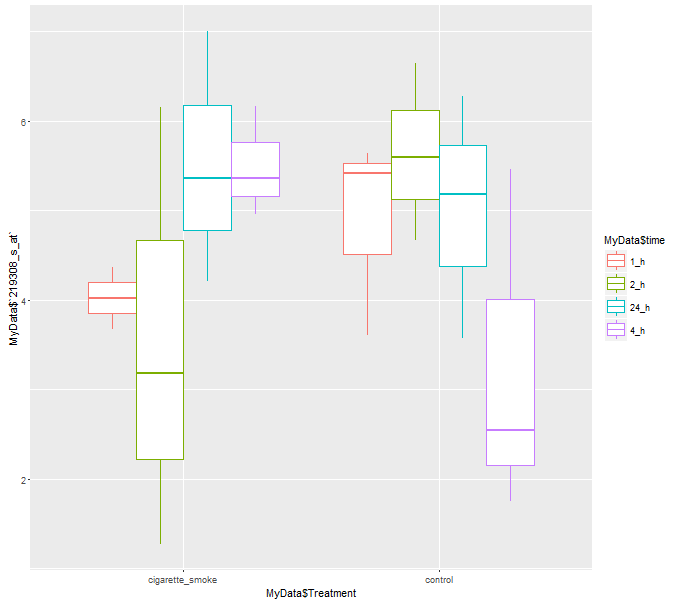
Note that the current variable time in this dataset is a factor (time column of design.txt). Create a new variable which is a numeric version of time (basically a new column in the metadata’s dataframe). It should represent a numeric count of hours. You will need this new quantitative variable later in the analysis.

**1.3 Basic graphing**

Create a plot showing the gene expression data for a random probe and the averages for all possible combinations of agent and time.

Hint: Choose a random probe first. Use position, panels or facets, and color to convey the treatment, time of the samples. Report the average expression of the selected probe for all possible combinations of treatment and time. Treat time like a factor in this question.

For this particular probe smoke of cigarette appeared to increase its expression over time.



**Question 2: Assessing data quality**

After loading the data, the second must-do thing is to sanity check the data.

**2.1 Examine the sample-to-sample correlations in a heatmap**

* Create the following sample-sample correlation matrices.
  + Order the samples by time; within each time group, sort on the other factor, i.e., agent (treatment).
  + Order the samples by agent; and within each agent group, sort by time.
* Interpret your results. What does the sample correlation matrix tell us about the overall impact of time and agent?

Looking at the heatmap of the matrix(attached pdf),there was no effect of the cigarette smoke in the cells. The color patterns in both control and cigarette smoke groups appeared to be same in the heat map.

**2.2 Assess the presence of outlier samples**

* Comment on the potential presence of outliers (if any) based on the sample correlation matrices created in Q2a.

There are 34 potential outliers

* Try to go beyond merely eyeballing, and make a quantitative statement about presence of an outlier: e.g., quantify for each sample, whether it `sticks out' compared to the other samples.
* If any sample does`stick out' from the previous step, examine it in the context of its experimental group (e.g., given a sample that is least similar to all other samples: does it correlate with samples treated with same agent better than other samples treated with a different agent?)

**Question 3: Differential expression with respect to treatment**

**3.1 Linear model**

* Fit a linear model, modeling expression level of each probe using treatment as a single covariate. (Hint: limma)
* Write out in English and as an equation the model you are using to assess differential expression. In the context of that model, what statistical test are you performing?

Equation: Y = Xα + ε, where Y is the response variable (gene expression), X is the treatment, α is the coefficients, and ε is the error rate from the fitting.

**3.2 Look at the hits**

* How many hits (probes) are associated with treatment at unadjusted p-value 1e-3? How many are associated with treatment at FDR 0.05?

By treatment (not discriminating by time), 43 genes were found to be differentially expressed at p.value 0.001 and with FDR 0.05 1238 genes.

* Take the top 50 probes as your `hits' and create a heatmap of their expression levels. Sort the hits by p-values and the samples by treatment.
* What is the (estimated) false discovery rate of this “hits” list? How many of these hits do we expect to be false discoveries?

**Question 4: Differential expression with respect to time**

For this question, time is treated as a quantitative covariate (unit: hours).

**4.1 Linear model**

You know the drill! Fit a linear model for assessing the effect of time on gene expression

* How many hits (probes) are associated with treatment at unadjusted p-value 1e-3?

How many are associated with treatment at FDR 0.05?

In this case we found 62 genes at p.value 0.001 and 951 genes at FDR 0.05

**Question 5: Differential expression analysis with a full model**

Like question 4, treat time as a quantitative variable in this question. We’re using both treatment and time as covariates in the full model.

**5.1 Quantify the number of hits for treatment**

* For how many probes is treatment a significant factor at the unadjusted p-value 1e-3, and at FDR 0.05 level?

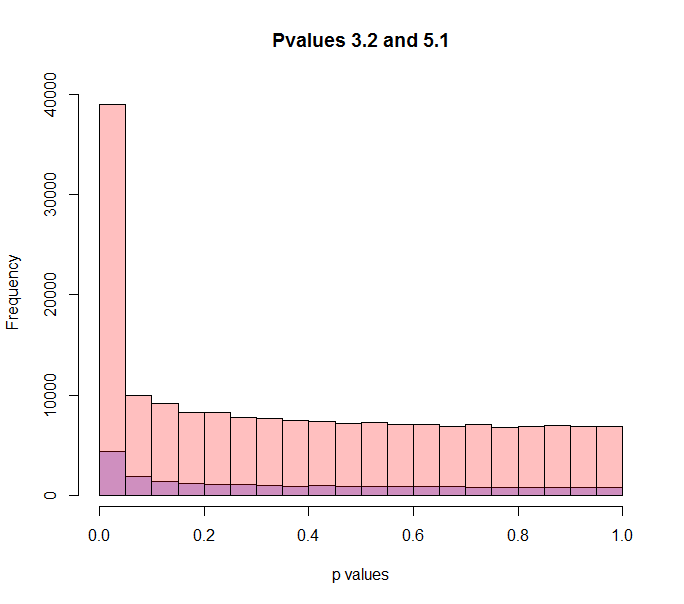
37 and 721 genes, respectively.

* Is this number different from what you reported in 3.2? Why? Quantify the proportion of overlapping probes among your hits, when using the unadjusted p-value threshold of 1e-3.

Yes, these numbers are different to the ones in 3.2 (43 and 1238 genes). Since the full model account for the two covariates (time and treatment) it is more stringent. In 3.2. false discoveries are influence by the comparison of samples at different time points.

* Plot the distributions of all the p-values for treatment when using both models, i.e., one from the model in Q3 and one from the full model in this question. Compare and comment on the similarity/differences in the shape of the distributions.

The figure below shows that although the distributions have similar shape, they differ in the frequencies, especially those of very low p-values. Thus, resulting in different amount of genes differentially expressed.



**5.2 Test the null hypothesis**

Null hypothesis: there’s no significant interaction between time and treatment.

* Explain in English what you are modeling with this interaction term (what does it represent?).

This means that for the different time points the two treatments have different expression patterns. For instance, an interaction time-treatment would be seen in the case of expression of probes increasing over time in the control and decrease in smoke\_cigarette group.

* For how many probes is the interaction effect significant at the unadjusted p-value 1e-3, and at FDR 0.05 level?

**5.3 Plot a few probes where the interaction does and does not matter**

Plot the expression levels for each sample on the y-axis and time on the x-axis. Color the points based on treatment group. Include the fitted regression lines in your plots.