Analysis of RNA Sequencing Data

RNA Sequencing is one of the most popular data acquisition technique for various labs. RNA is a very important molecule which decodes life itself. Its measurement can tell us a lot about processes going inside a cell. We have one such measurement for T-Cells at different time points in its cycle of activation. The measurement consists of ~24000 genes for the different samples taken at different time points and different conditions.

Distil the multidimensional information to dimensions where simple biologists can comprehend it

A simple way to do it is to perform PCA on the data and make sense with already existing metadata. There is, however, one small hurdle in doing this, RNA Sequencing produces data with a negative binomial distribution. One would need to normalize the data before doing the PCA.

Hints: DESeq2, pcaExplorer

Questions:

1. Does the PCA changes drastically if we perform PCA on the top 500 (in terms of variance) as compared to top 5000 genes?
2. If the answer is yes, then would we consider the system (the cell) to be noisy? If the above answer is no, then what can we say about the system as a whole?
3. What is the overall takeaway from PCA plots and its correlation with metadata?

Remove noise due to experimental conditions

There is a very high probability that experiments performed on different days and different labs will produce a different count of RNA molecules. This is known as the batch effect and it is very important to remove it to uncover the real biology. The above data consists of two experiments and there is a batch effect among them. One way to do this is to model the batch effect and remove it from the data.

Questions:

1. Is the batch effect consistent amongst all the donors in the experiment?
2. Does overall biology make sense after removing the batch effect? What has changed after batch effect removal?

Hints: SVA, EdgeR

Tracking the analysis

A very important part of a analysis is its reproducibility. To make analysis reproducible we need to preserve threads of thoughts and share it. One way to do this is to use a framework which preserves all the steps in an analysis and what else can be better than git. Fortunately, we already have solutions such as knowledge repo available.

Task: Use knowledge repo to track the analysis done above and host it on GitHub (don't make the data public)

Share the analysis

While we are doing the analysis, we would like to share it with our colleagues and ask for their opinion.

Task: Host the knowledge repo on a server (Heroku maybe) and configure it to update itself after for every commit (real time isn't necessary here, a cron job would do)

Bonus Question

Public data is a big source of data in biology. Multiple efforts by multiple agencies around the world are working on a very specific problem in biology (TCGA, GEO, LINCS Project etc). Interpreting the data and doing an integrative analysis is a daunting and challenging task, an example this could be, take the data from TCGA, find subtypes within particular cancer, confirm those with GEO and target those subtypes with help of LINCS, this process might take weeks or even months to give some fruitful results. Having a machine do this task will be immensely powerful. One such task that is relevant to the data above is to find all the mouse and human studies on T cell activation and validate our findings with those studies.

Questions:

1. How would you find the studies which talk about human and mouse T cell activation? (A general workflow which can be automated will be helpful)
2. How would one represent the data which give a sense that the data above agrees with the public data? (One such figure is mandatory, and an outline of dashboard would be helpful)