Genetics & Genomics – Genomics Project

In this exercise we are working with a transcriptomics dataset containing 80 samples. 40 samples are biological replicates of human preadipocytes, and the 40 others are differentiated adipocytes. We aim at visualizing these two conditions and identify the key genes that are markers of adipocytes.

The main dataset is a simple "count matrix" containing genes (in rows) and samples (in columns).

1. Load and visualize the data

a. Start by loading the count matrix (*transcriptomics.dataset.txt*) and display it. Genes are represented by their Ensembl IDs, and should be the row indexes. What is the size of the matrix?

b. What type of data do you have? Normalized or raw read counts (integer)?

c. Identify the two groups of samples using the column names. Create a dictionary with mapping your sample names to their groups (Adipo, PreAdipo).

d. Filtering: Remove the genes that are not expressed in any sample (i.e. only 0s)

e. Normalization: Check the sum of the reads/counts per sample, to see if you need to normalize for library depth. Normalize if needed, using a CPM calculation. Show the depth as a barplot.

f. Plot the PCA and color by condition.

g. Transformation: Perform log2(1+x) transformation of your data.

h. Plot the PCA (of the logged data) and color by condition. Which PCA looks more like what you would expect? Continue the next questions using the more appropriate Normalization/Transformation of your dataset.

2. Differential gene expression

In this part, we will create a separate pandas DataFrame "de_results" to store all results. Rows will be the filtered genes, and columns will be pval / fdr / log2_mean_preadipo / log2_mean_adipo / log2_fold_change, that will be computed in the following questions:

a. Perform a **two-tailed independent t-test** on every gene to check if it's differentially expressed between the two groups (preadipo vs adipo). Store its pvalues in the "de_results" DataFrame.

   Note: You can use the *scipy.stats* package. Note as well that here the two populations (adipo and preadipo) may not have equal population variance, so maybe a Welch's t-test (instead of a standard Student's t-test) would be more adapted.

b. Adjust the p-values for multiple testing (use FDR correction), and store the output in the fdr column of the "de_results" dataframe. Why do we need to do that?

   Note: You can use the *statsmodels.stats.multitest* package

c. How many genes are differentially expressed (DE) based on this calculation? (select a threshold, for e.g. FDR < 5%)

d. Here we want to focus on genes that are up-regulated in the adipocytes (marker genes). For selecting these, we will first compute the arithmetic mean of gene expression in each group, for each gene, and put the results (logged mean) in the columns log2_mean_preadipo and log2_mean_adipo of the de_results DataFrame.

   Note: Be careful if your expression matrix is logged , not to compute the logarithmic mean

e. Now compute the log2_fold_change between the two groups.

f. Filter the de_results DataFrame and keep only genes that have FDR < 5% and abs(FC) > 2 (i.e. log2_fold_change > 1). Sort by fold-change.

3. Investigating top marker genes

a. Use the gene annotation file (*Homo_sapiens.GRCh37.75.gene_annotation.txt*) to annotate your Ensembl IDs. Annotate the genes in the filtered de_results DataFrame by adding a 'gene_symbol' column.

b. Plot the expression of the top marker gene as two side-by-side plots, one with expression as a gradient color in the PCA, and another as two boxplots showing expression of this gene for each sample in the two groups. Put the Ensembl name and Gene symbol in the title of the plot.

c. Amongst the most common marker genes for adipocytes are Adipoq / Fabp4 / PParg. What is their rank in your list of DE genes? Plot their expression the same way you did for **3.b**.

   Note: If possible, avoid repetition of similar code

d. Plot the top 20 up-regulated genes and top 20 down-regulated genes as a heatmap. Try to make the most "publishable" figure. For e.g. by displaying the dendrograms, annotating the samples, pick a nice colormap, etc.

   Note: You can use the *clustermap* function in the *seaborn* package

4. Functional enrichment (Bonus)

When we perform differential expression, we may end up with a big list of genes, not knowing their biological function overall. An easy way to get some biological meaning of a list of genes is to perform functional enrichment into an annotated database. These databases contain annotation of genes into certain cell types, biological pathways, etc… Most of the time it is represented as gene sets, i.e. and array of genes that are annotated as "Adipogenesis" for e.g. So, the database contains plenty of these genesets that annotate cell types/pathways/etc..

When you have a list of genes (for e.g. differentially expressed genes), you can find out if they are enriched in a specific gene set/category by performing a Fisher's Exact Test on the following contingency table:

	D	$\bar{D}$	Sum
C	1962	9803	11765
$\bar{C}$	118	709	827
Sum	2080	10512	12592

C: in category;
$\bar{C}$: not in category;
D: DE genes;
$\bar{D}$: non-DE genes.
doi:10.1371/journal.pone.0046128.t001

a. Create the 2x2 contingency table of this example (as a matrix with 2 rows/2columns, the sums are not needed)

b. Calculate the p-value associated, is it significant at 5% threshold?

c. Download the “Human_Gene_Atlas” gene set database from https://amp.pharm.mssm.edu/Enrichr/#stats. This file contains marker genes for many cell types in Human. You will see that the file format is very peculiar, so find a way to read the file in Python and store the gene sets.

   Note: Keep only genes that overlap with the genes in your dataset. All others could not be find as DE anyways, so they should not be counted.

d. Perform functional enrichment on the Adipocyte gene set, with your top 20 list of up-regulated DE genes in the adipocyte samples. Is it significant?

e. Now, run the enrichment on each gene set of the database. Store the p-values and correct for multiple testing. Then sort the results by p-values and check the significant genesets at FDR5%. Which gene sets are the most significant? Does it make sense?