Home assignment 1:
Exploratory analysis and visualization

Jose Francisco Sánchez Herrero

Deadline: November 14th, 2024

Introduction to the topic and dataset

You are a team of bioinformaticians working for a company that provides scientific, statistical and bioinformatic consulting services. A customer from a very important sherry wine trade company wants to understand the differences between strains of *Saccharomyces cerevisiae* for different sherry wines products that they elaborate.

As you may know, flour strains of this yeast are principal microbial agents responsible for biological wine aging used for production of wines. The flour yeast velum formed on the surface of fortified fermented must is a major adaptive and technological characteristic of flour yeasts that helps them to withstanding stressful winemaking conditions and ensures specific biochemical and sensory oxidative alterations typical for sherry wines.

The user produces **two types of sherry wines (A and B)**, with different flavors, alcohol percentage and color. They are interested in deciphering if the differences between sherry wines classes are due to genetic differences in yeast or any other biochemical process that has not been controlled.

We recommended applying RNAseq technology for transcriptome analysis of these two industrial flour yeasts strains at different steps of velum development (**6 different times**) and multiple replicates (**4 replicates**). We ended up with 48 samples analyzed.

RNAseq reads were automatically quality checked, trimmed and filtered following default settings of packages such as fastqc and trimmomatic included in a RNA seq bioinformatics pipeline (https://github.com/HCGB-IGTP/RNA_seq_pipe).

During the quality check, a sample produced unusual results but was not discarded. Please take into account this and discard it if necessary, after the normalization process.

Mapping to the reference genome and feature count of genes was produced using Hisat2 and featureCounts packages, also included in the previous pipeline mentioned.

For each of the 4 replicates, a gene count matrix was generated and saved as a comma separated file within the data folder provided.

Your goal in this project is to decipher if there are genetic differences between sherry wine yeast strains that might be reflected in the wine tasting.

I encourage you to show what you know and interpret what you see in a plot using the theory we learned.

Provide code and answers for the following questions and add different visualization and/or code alternatives, additional items that might be useful (reviewed in this or other lessons), everything will be taken into consideration.

- **QUESTION #1: Load, adapt the data and create metadata**

Load the four different replicates provided as separate files into separate variables. Check data is correctly loaded, and all samples have the same dimensions.

Remember to adapt the data accordingly. We need columns of the tables to represent variables (genes) and rows to be samples. Also, we need to merge all tables of observations. Check all samples are in the same order.

Create a **meta data dataframe** containing information regarding replicate, strains and time points. Information is either included in the sample names or in the introduction to the dataset stated above.

- **QUESTION #2: PCA representation**

Represent the data with a PCA projection: non-scaled, scaled, normalized.

Argue whether batch effects are present between the 4 replicates.

Remove outliers if necessary.

Show the different steps in the process and comparison with raw-data and normalized.

Scale or normalize data appropriately.

Identify and show any correlation with expression if necessary.

- **QUESTION #3: tSNE representation**

Represent the data with a tSNE projection: raw data and normalized.

Again, argue whether batch effects are present between the 4 replicates.

Remove outliers if necessary. Scale or normalize data appropriately.

Show the different steps in the process and comparison with raw-data and normalized.

- **QUESTION #4: tSNE parameters**

Test the effect of reproducibility, perplexity and iterations. Use normalized dataset only.

Provide comments and thoughts about it.

- **QUESTION #5: Final interpretation**

Create final representation for PCA and tSNE.

Using the normalized dataset get a conclusion and comments for PC1 and PC2.

Provide comments and thoughts about it. Do you get to the same conclusions?