A PRIMER ON DATA VISUALIZATION IN BIOINFORMATICS

Finding COVID-related genes as a guiding example

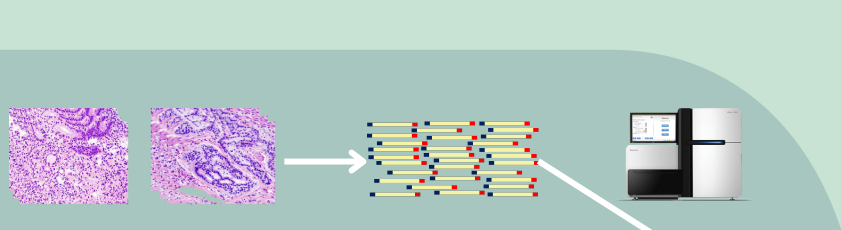
Biological information is contained in genes. Genes (DNA) express themselves by transcribing their sequence into RNA fragments (transcripts). Knowing which genes express differently between healthy people and COVID patients can help us to understand and cure the disease.

We introduce some visualizations used in the search for those genes. This is done by sequencing blood samples and comparing the abundance of RNA transcripts in each group: "more expression, more abundance".

How can we measure transcript abundance?

INPUT DATA

From each blood sample, RNA is extracted and prepared to be sequenced, a process that yields a large number of copies of each RNA fragment.
Sequenced transcripts are aligned to a reference genome to know which gene is expressing each transcript. It is like reconstructing a puzzle.
Gene expression is quantified by counting the number of fragments aligned with each gene.
This is summarized in the counts' matrix, with one row per gene and one column per sample.

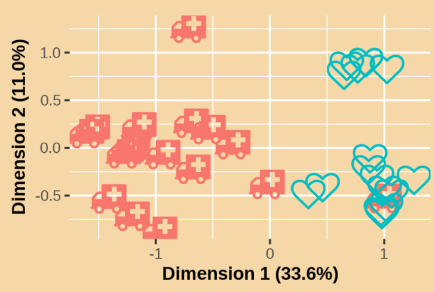


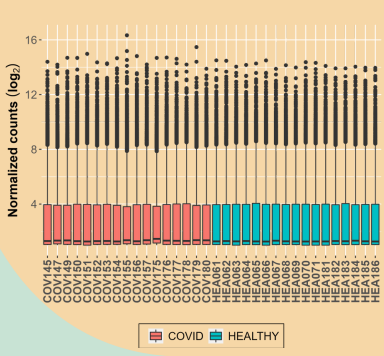
	Healthy 1	Healthy 2	Healthy 3	COVID 1	COVID 2	COVID 3
Gene A	6	5	7	17	10	11
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

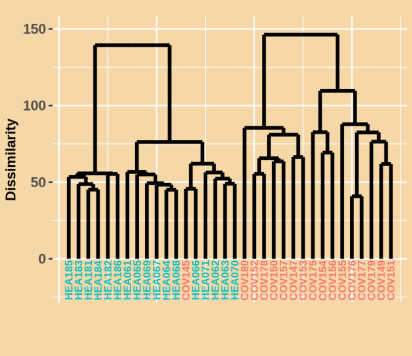
Gene C is a potential candidate to be called under-expressed in COVID patients

Has there been any problem at generating the data?

QUALITY CONTROL & EXPLORATION







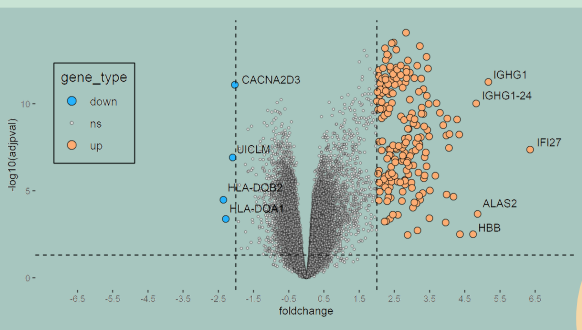
Boxplots and other graphics help to check data distributions.
Ideally, one might expect that samples tend to be more similar within groups than between groups. Distinct techniques such as PCA or Hierarchical clustering are used to check this assumption.

PCA and dendrogram suggest that there is a COVID patient similar to HEALTHY ones

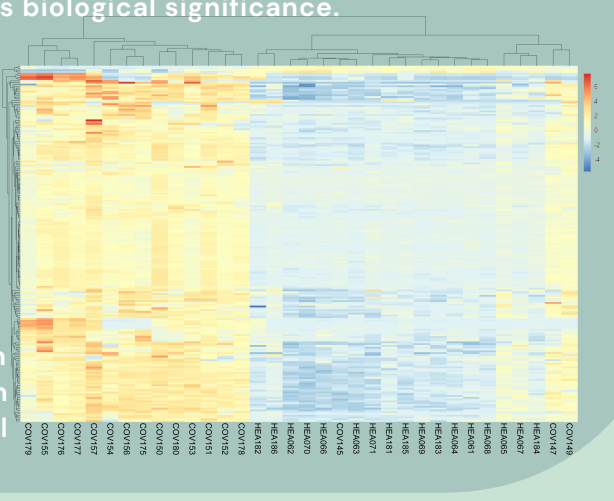
Which genes express differently in the two groups?

DATA ANALYSIS

A statistical test allows selecting significant differentially expressed genes. The volcano plot shows statistical versus biological significance.

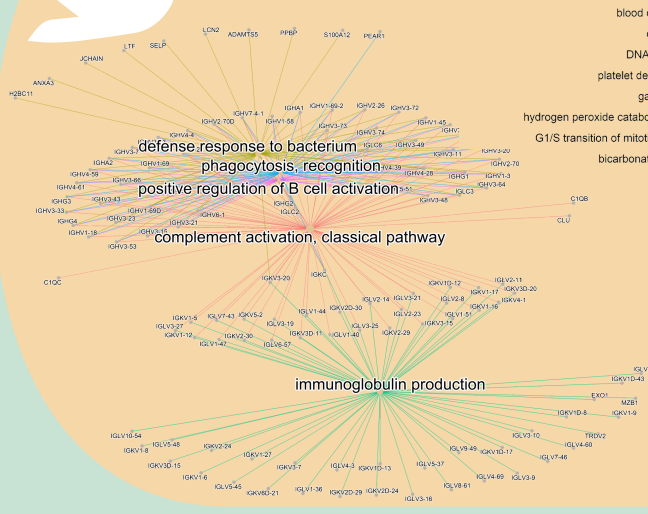


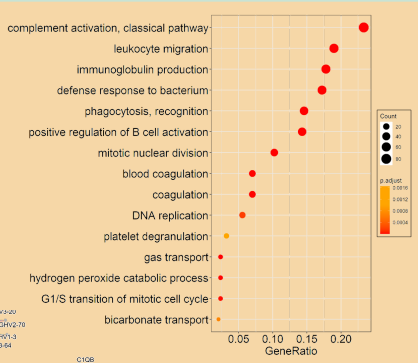
Most differentially expressed genes are over-expressed in COVID patients



Heatmap displays the expressions of selected genes in a grid (genes in rows & samples in columns). The color scale reflects the intensity of gene expression in each sample. On the margins, dendrograms group genes or samples based on the similarity of their gene expression pattern. This is useful for identifying genes that are commonly regulated.

What does that list mean biologically?





The dotplot shows the 15 categories best represented by our COVID-related genes

BIOLOGICAL INTERPRETATION

Genes are *annotated* in different knowledge databases by terms or categories describing their biological role.
The distribution of annotations of selected genes is compared with the distribution of the same annotations in the genome. This allows determining which biological processes might be associated with our gene list.
We end up linking those differentially expressed genes that are included in the most represented biological categories using a network plot (bottom left-hand figure)

Analyses based on data from: ***Arunachalam et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science. 2020 Sep 4;369(6508):1210-20. doi: 10.1126/science.abc6261.***

UNIVERSITAT DE BARCELONA
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
GRBIO

Code of analysis available at: https://github.com/GRBio