OMICS MEET ML

Things to take care of

Pietro Franceschi

Unit of Computational Biology - Fondazione E. Mach

2022-02-21

My Scientific Id

- **PhD** in Physics
- **Research Interests**: analysis of complex data, method development, metabolomics, mass spectrometry
- **Tools**: R and (less often) Python
- **Other Interests**: clarinet playing, bonsai, aquarium, DIY in general, …

Omics

What are omics

Adding the suffix *omic* to a word is a way to indicate the **desire** of performing an **holistic large scale investigation** of a specific subject

- genomics (and meta-genomics)
- proteomics
- metabolomics (with its extensions like lipidomics, gl ycomics, …)
- …

The rise of *quantitative* technologies are transforming almost all disciplines into *omics* … tholinomics, petroleomics

Common Ideas

- challenging
- comprehensive and holistic (as much as possible)
- data rich (measuring a large bunch of variables)
- complex (data processing and interpretation require bioinformatics)
- multidisciplinary (nobody can do everything alone)

The role of bioinformatics/biostatitics

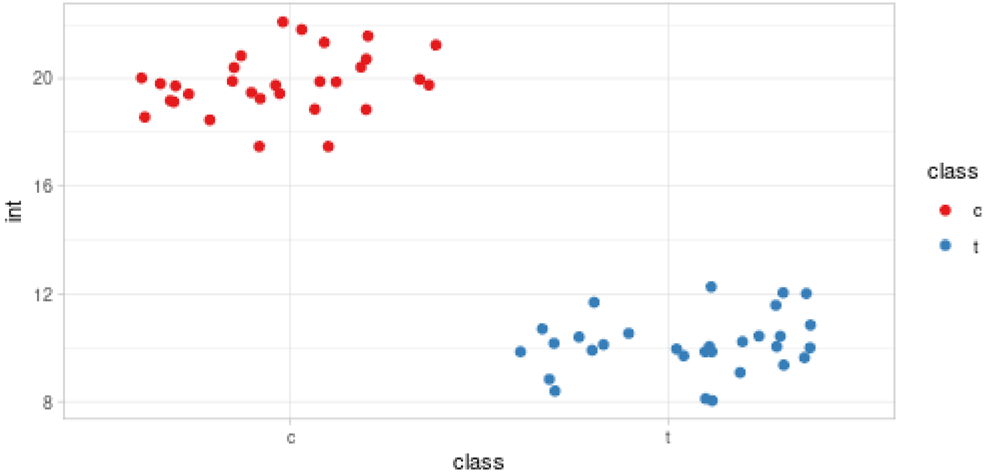
Spot/show a result present in data

Show/assess that my result holds for all the population … **A scientific result have to be general**!

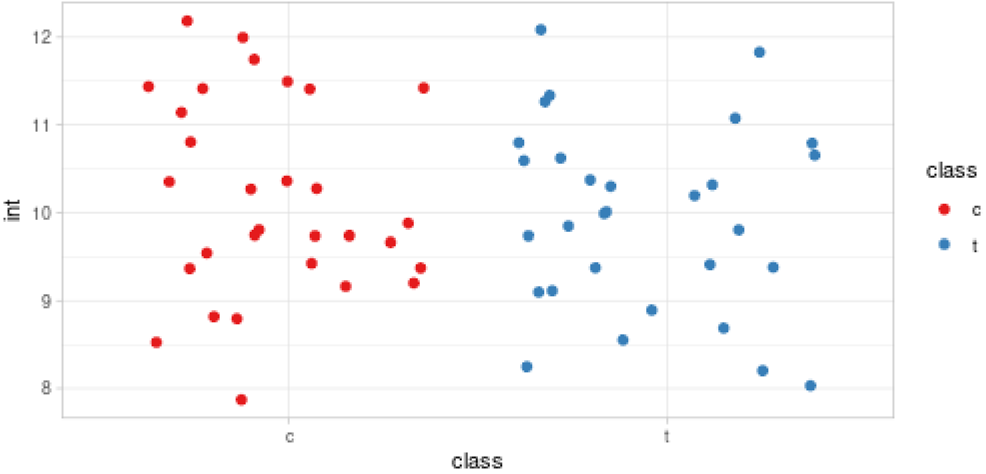
Allow me to reproduce my results

What is a **result**? A result is **organization** !

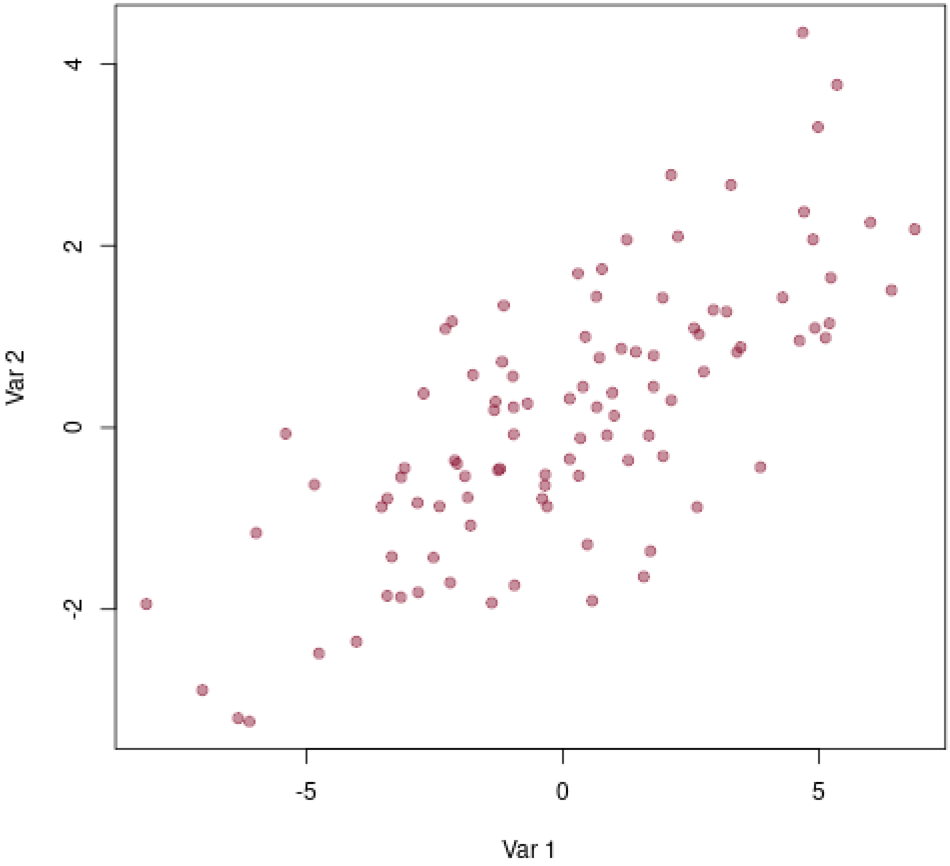
A biomarker! :-)



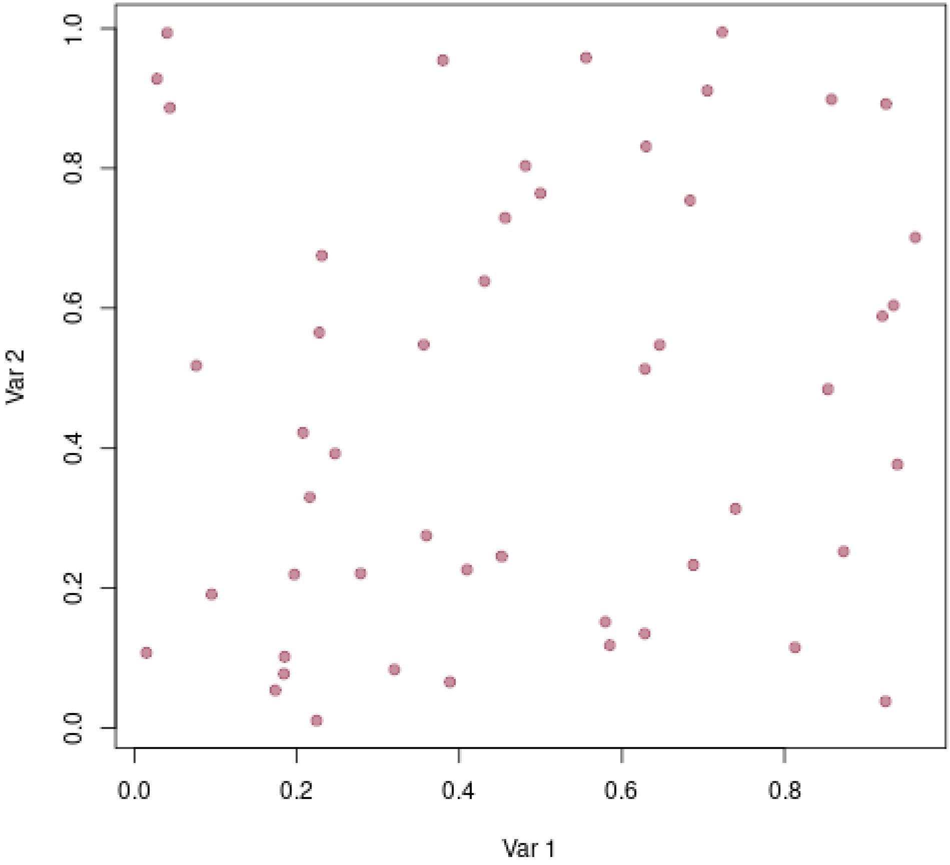
A non-biomarker! :-(



Correlated variables! :-)



Uncorrelated variables ... :-(



Data analysis goals

- Visualize and spot these results in my data matrix (**Exploratory Data Analysis**)
- Generalize: at least *give a confidence* to my desire of generality (**statistical analysis and ML**)

Data deluge

- untargeted metabolomics: 10000 variables per sample
- targeted metabolomics: 300 metabolites per sample
- proteomics: 3000 proteins per sample
- metagenomics: 100000 OTUs
- NGS: many, many
- phenotyping with sensors (10 variables, every day/every half an hour)

Data matrices

Analogous of an Excel table

- rows are samples
- columns are variables

By definition, in *omics* the number of variables largely exceed the number of samples, and *technological development worsen* this unbalancing

Characteristics of omics data you should always remember

Sample to variable unbalancing

In a typical *omic* experiment the number of variables you measure largely exceeds the number of samples

It is **not** unlikely that the organization you measure is there only by chance

This is the result of **sampling**

As we will see, the chance of finding random organization grows with the number of variables we measure

Presence of unknown sub populations

The omic technology you are using to investigate your population will be able to discover **unexpected** and **hidden** structure of your sample

You are looking to your samples with a sort of *augmented reality* device

Uniform groups are not anymore uniform !

Multi level experimental designs

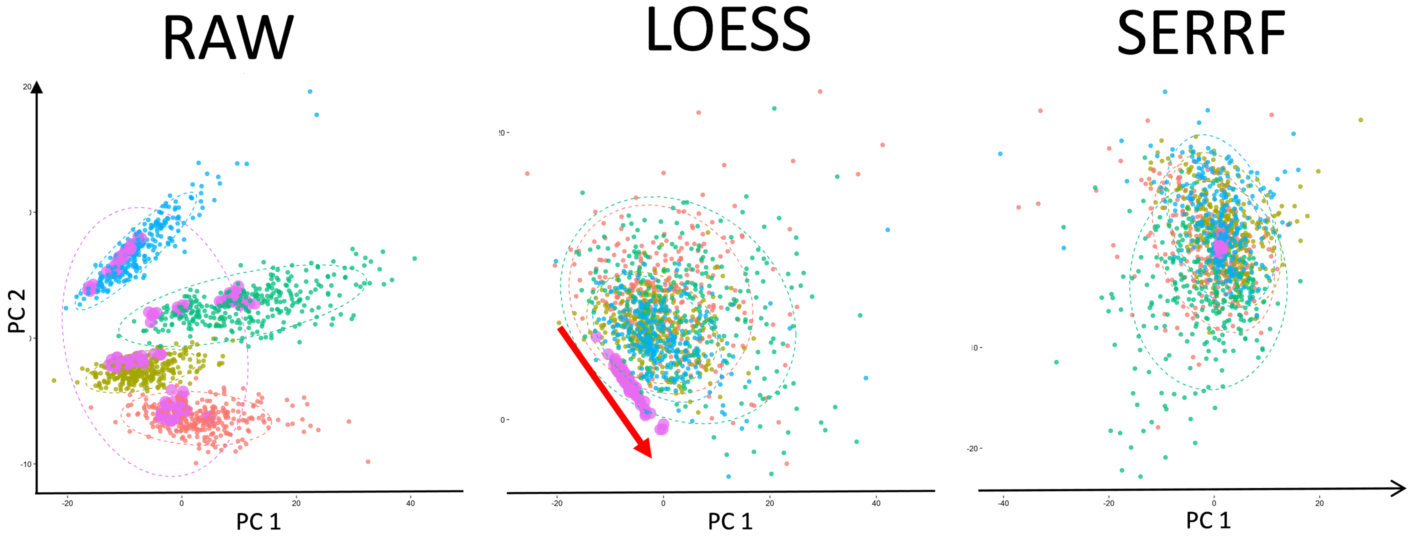
With the increasing availability of low cost *omic* technologies, we are able to apply such investigations to **multilevel experimental designs**

Samples are not independent!

Question

How can I take this aspects into account when analyzing my data?

Batch effects



RAW

LOESS

SERRF

https://slfan2013.github.io/SERRF-online/#

High Dynamic Range

A technology with **high dynamic range** is able to measure quantities over a large range of intensities/abundances

In other words: you measure together things that are abundant and things that are rare

E.g. In metabolomics concentrations can vary over 6 order of magnitude

Questions

Should we scale? Is reliability an issue?

Missing Values

Missing values are holes in my data matrix ... remember that **0 is not missing! Zero is zero**

They arise

Errors (somewhere)

Low intensities/abundances

Beware!

If we "fill in" missing values with the wrong data we bias our analysis

Missing Values: questions

Should I always fill them?

What number should I put there?

How can I be sure that my choice was good?