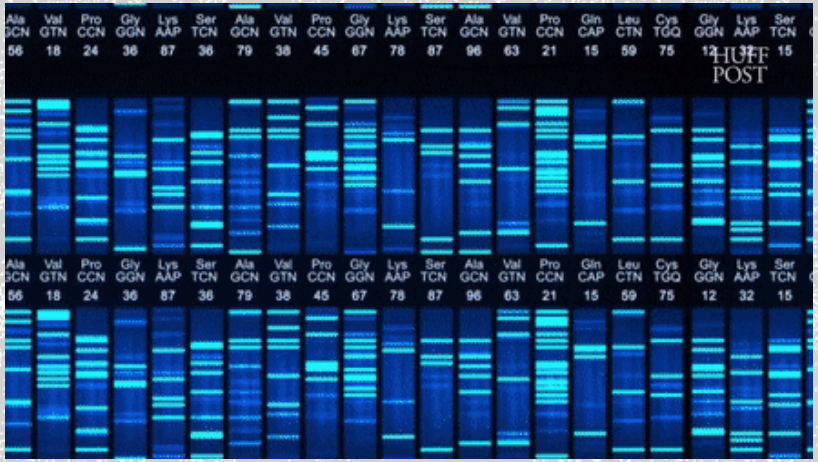
OMIC DATA



Descriptive analysis and differential expression

José Manuel Sánchez Santos
Dpto. Estadística - CIC-Usal

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

CiC
CENTRO DE INVESTIGACIÓN DEL CÁNCER

DATA FRAME

Multiple data frames

Variables \ Samples	S_1	S_2	…	S_m
Gene/Protein 1	X_{11}	X_{12}	…	X_{1m}
Gene/Protein 2	X_{21}	X_{22}	…	X_{2m}
…	…	…	…	…
Gene/Protein n	X_{n1}	X_{n2}	…	X_{nm}
Factor 1	F_{11}	F_{12}	…	F_{1m}
…	…	…	…	…
Factor k	F_{k1}	F_{k2}	…	F_{km}

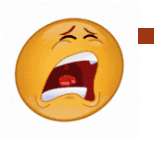
Omic data (gene/protein expression from microarrays, RNASeq,…)
-Genomic data
-Transcriptomic data

Phenotypic data (genre, age groups, patient type, cell type, cancer type, experiments, batch…)

OMIC DATA

- Gene expression values up to 2^{20}, then log-scale.

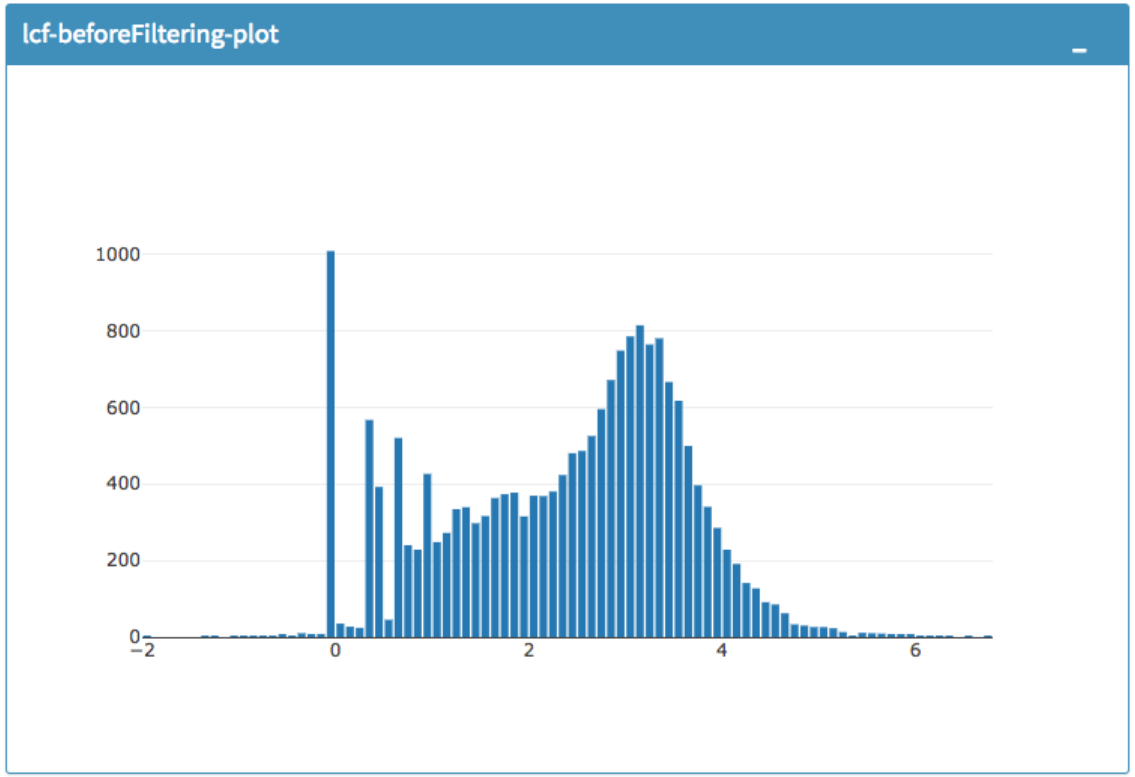
- Lots of genes with low expression.

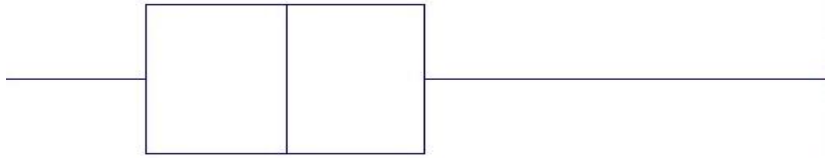


- Very few genes with high expression.

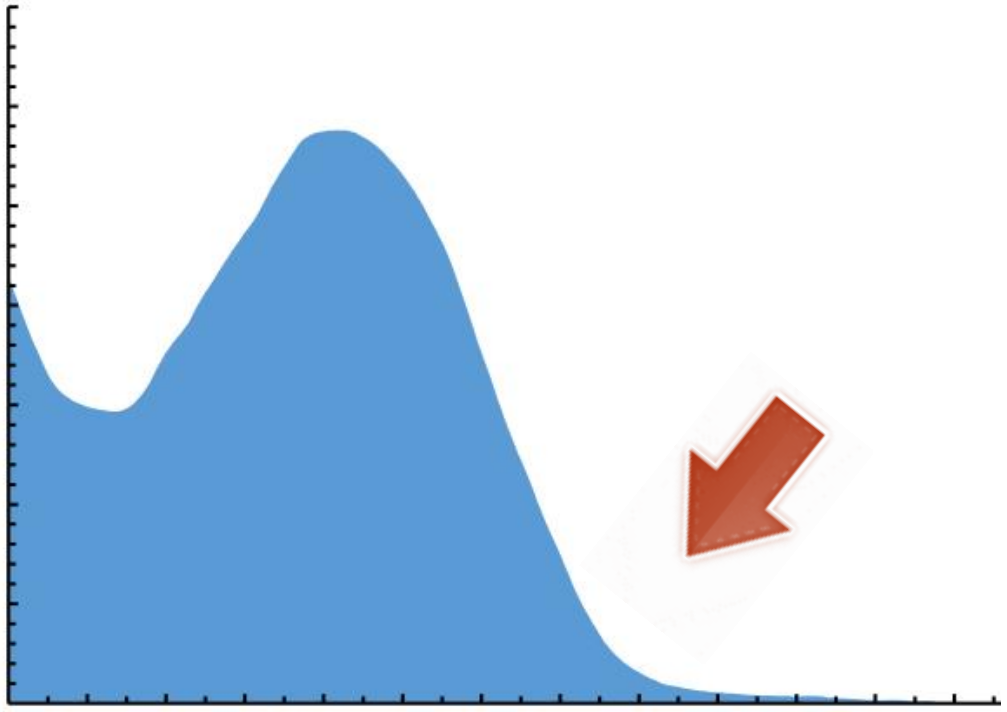
- Expression is not Normal (Gaussian) distributed.

- Plots: histogram, density, boxplot, MA-plot…

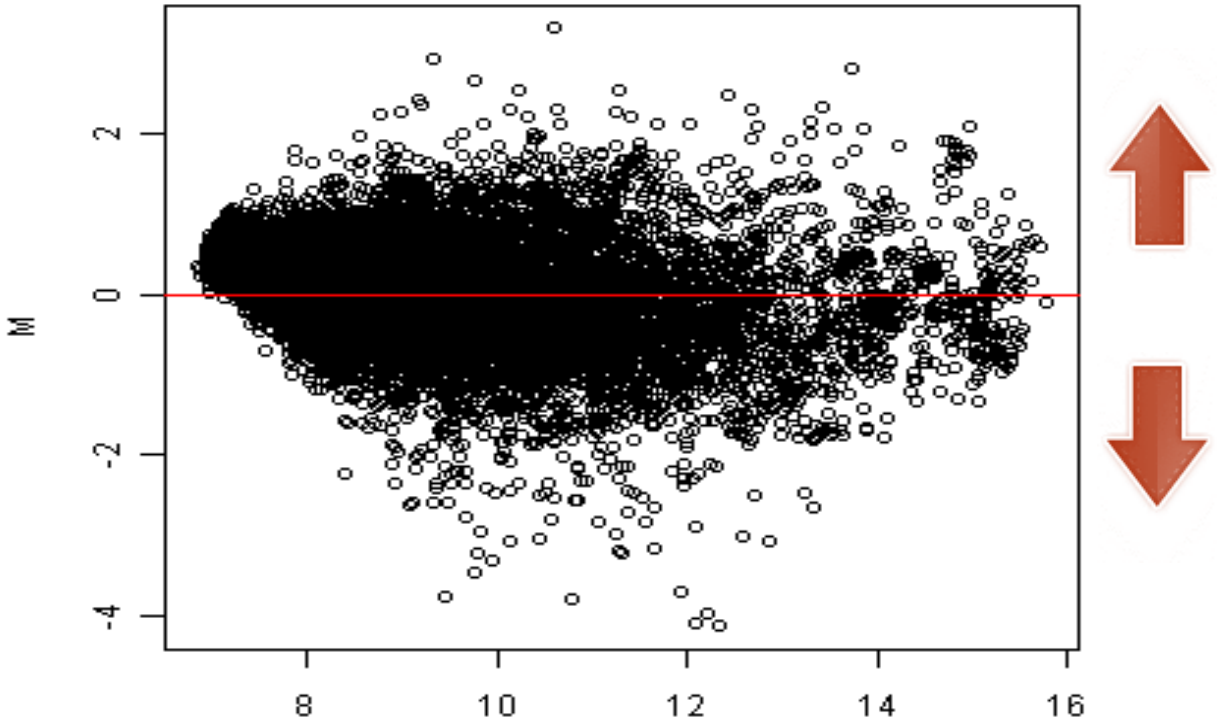




OMIC DATA



One microarray density plot

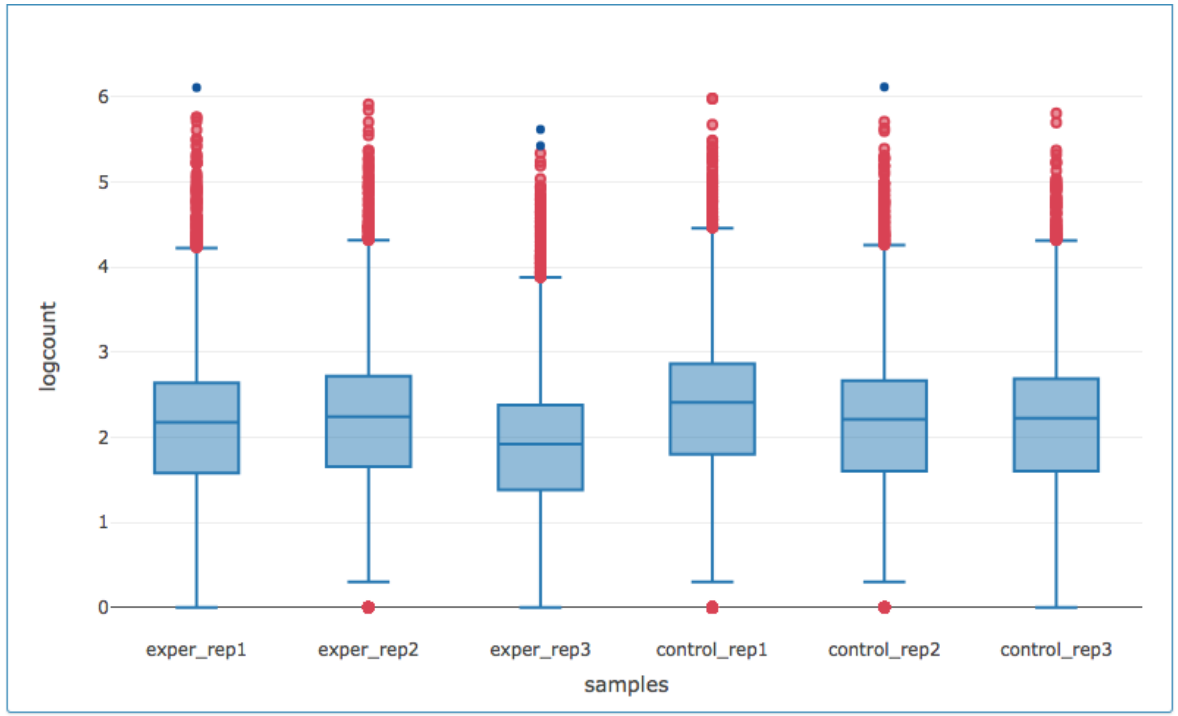


One microarray MA-plot (expresión vs mean)

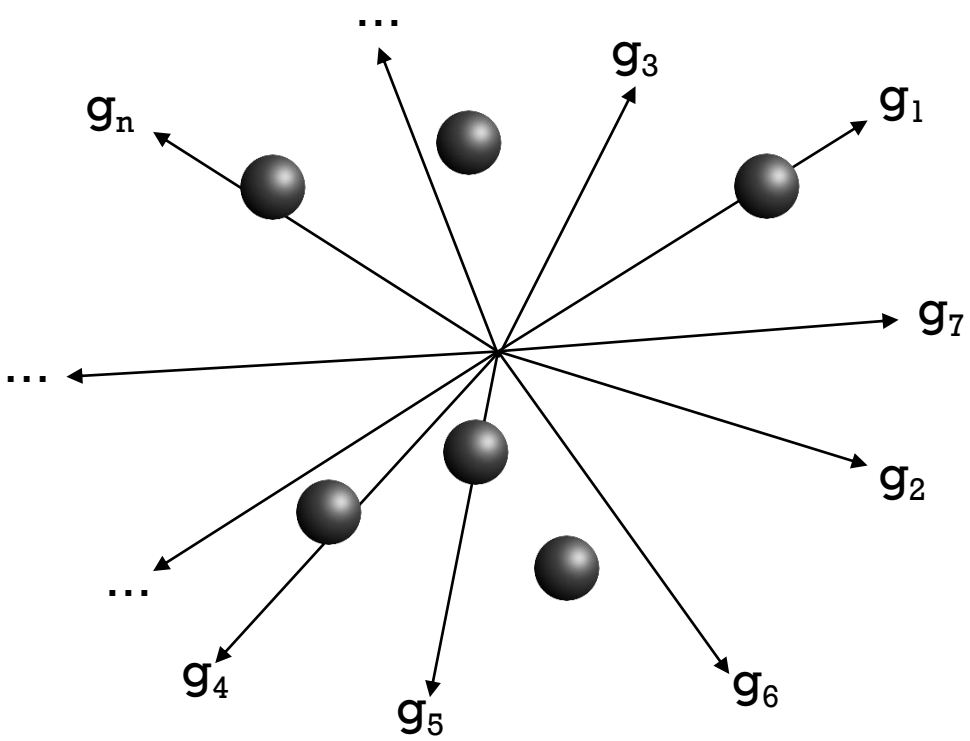
PHENOTYPIC DATA

WARNING

- Biological variability.
- Different experimental conditions.
- Samples from different sources.
- Join datasets from different experiments or batches.
- Plots: joint histograms / density / boxplots, PCA plot.

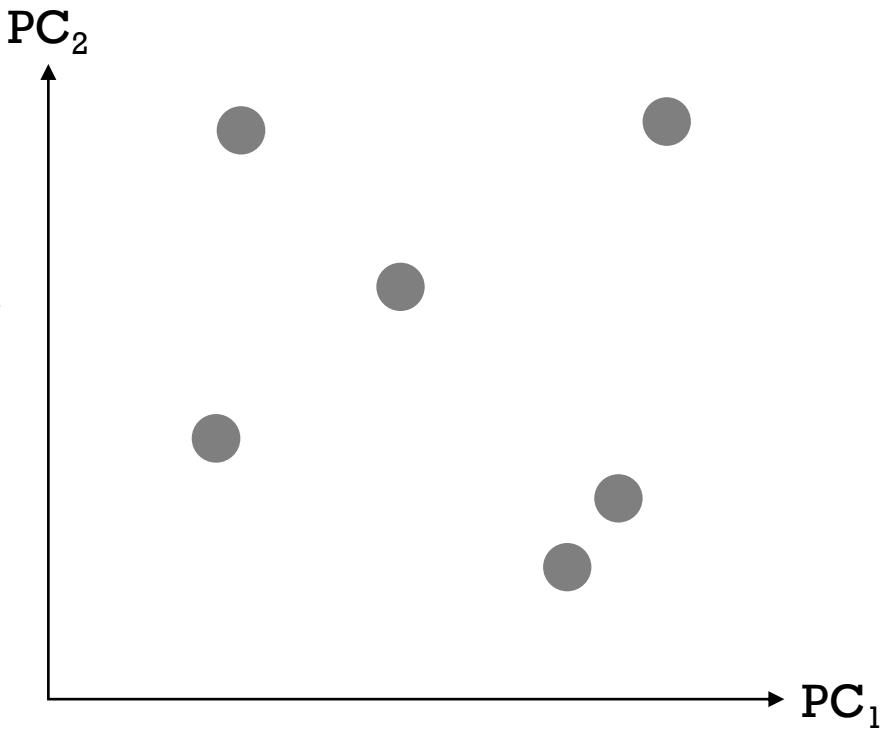


PHENOTYPIC DATA



(imagine) n dimensional space, each ● is a sample

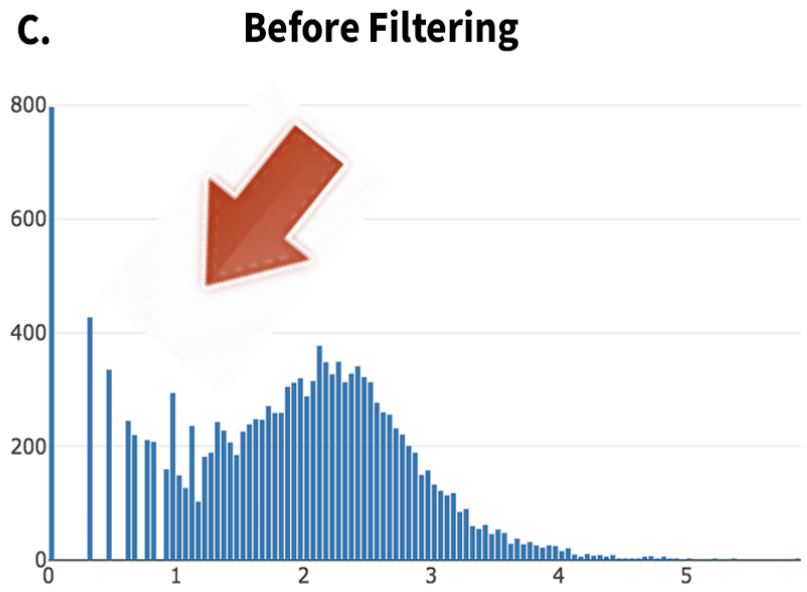
reduce

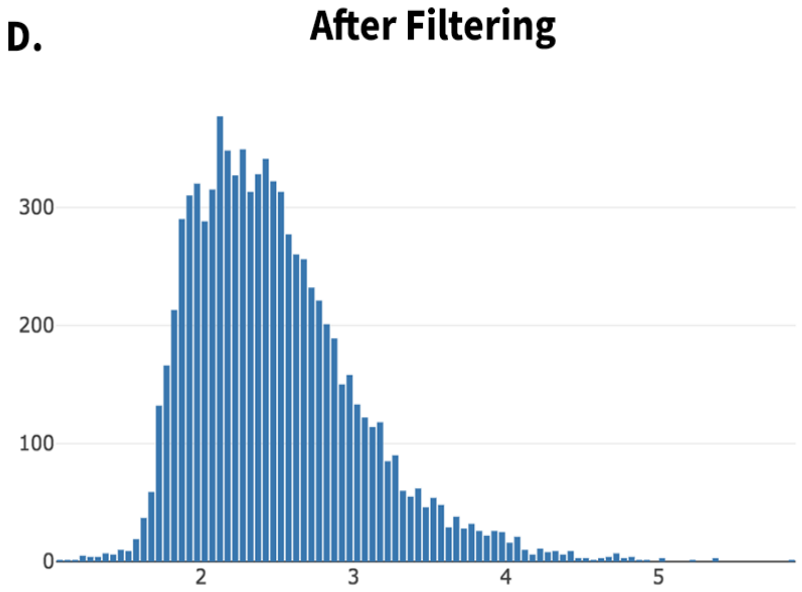


PCA: n dimensional space but plot of 2 axis explaining most variability

QUALITY CONTROL AND CORRECTIONS

Usually remove low expressions (low count filtering, background correction…)

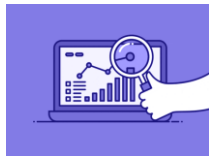


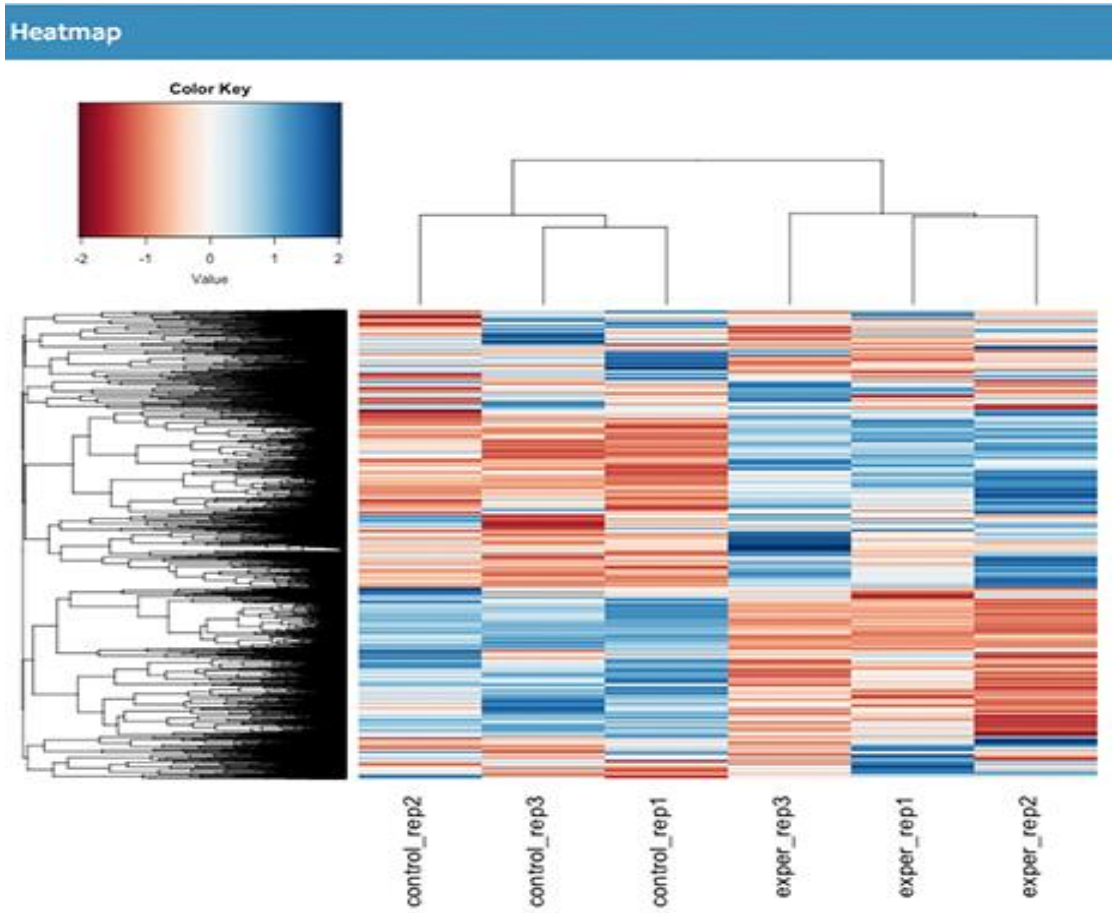


QUALITY CONTROL AND CORRECTIONS

Heatmap:

- Search patterns.
- Search batch effects.



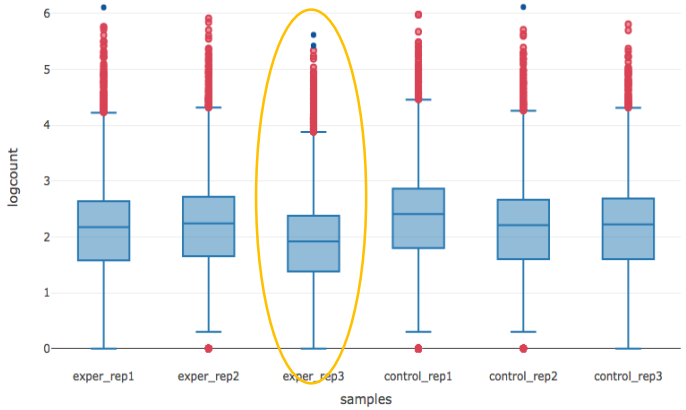


QUALITY CONTROL AND CORRECTIONS

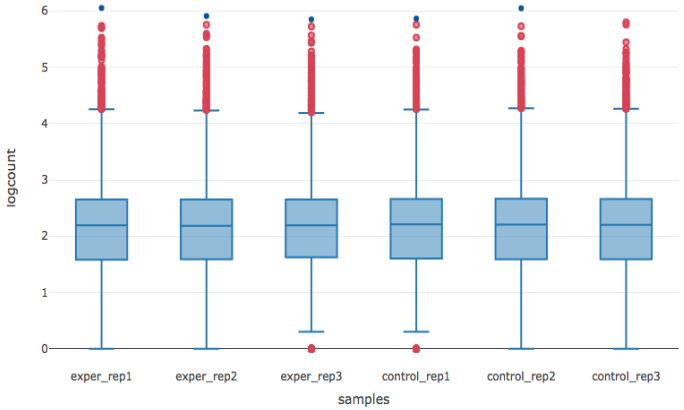
Normalize:

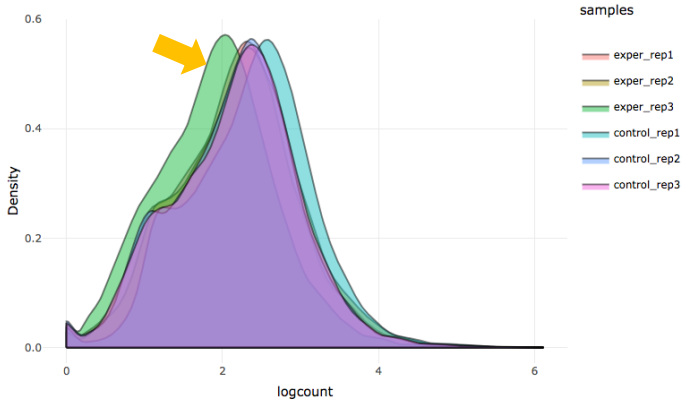
- Make the expression signal comparable.
- RMA, MAS5, fRMA…

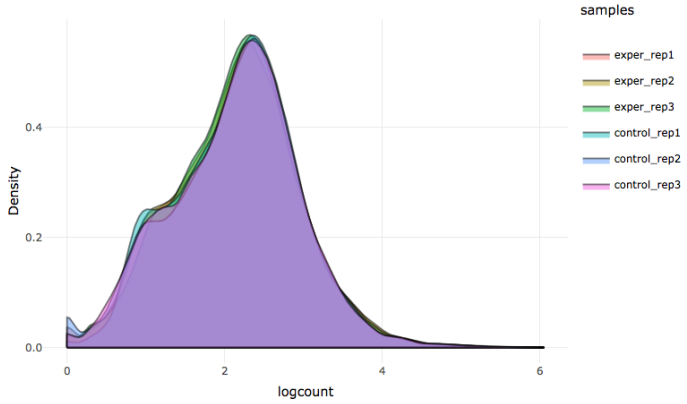
Before normalization



After normalization





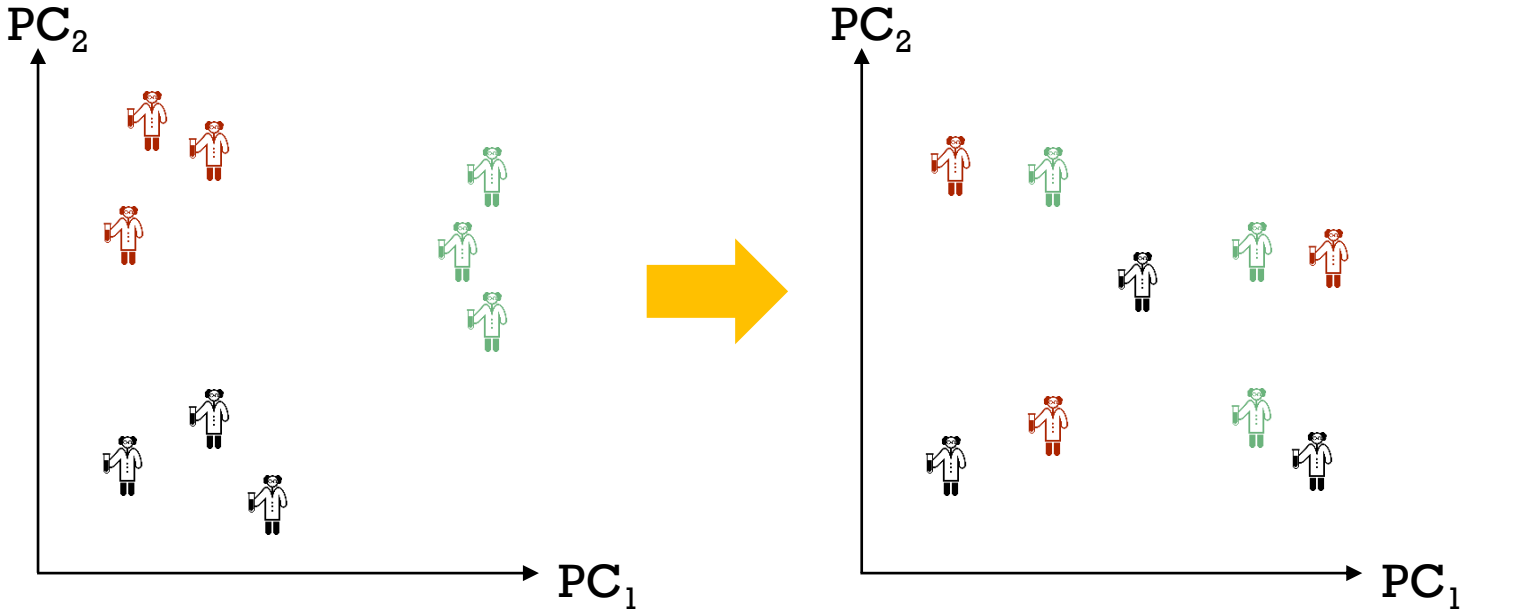


QUALITY CONTROL AND CORRECTIONS

Batch effects correction:

- Correct the technical sources of variation, such as different processing times or different handlers.

- Combat, f-RMA, TMM, RLE, upperquartile…



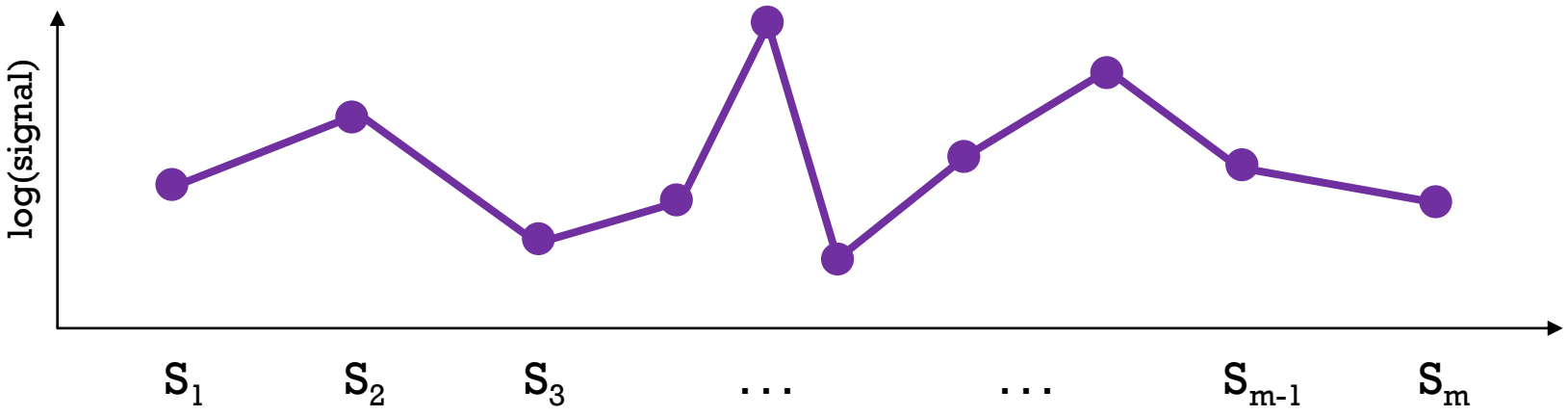
PCA of samples with batch effect

PCA of samples without batch effect

DIFFERENTIAL EXPRESSION

Gene/Protein expression profile: expression signal in each sample

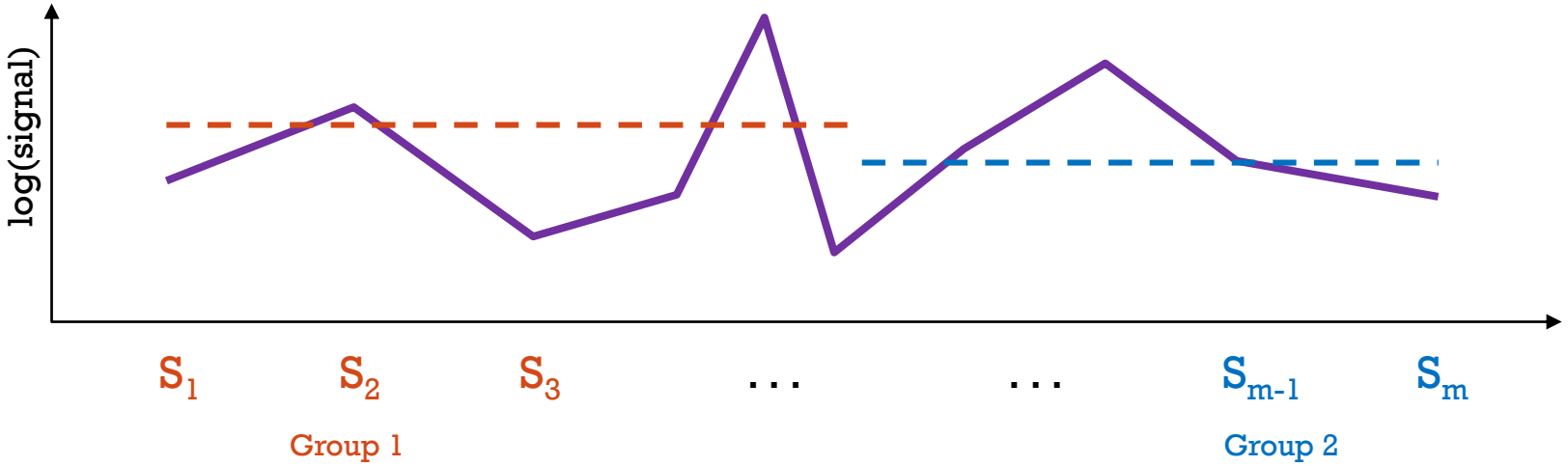
	S_1	S_2	S_3	…	S_{m-1}	S_m
Gene/Protein	X_{11}	X_{12}	X_{13}	…	$X_{1(m-1)}$	X_{1m}



DIFFERENTIAL EXPRESSION

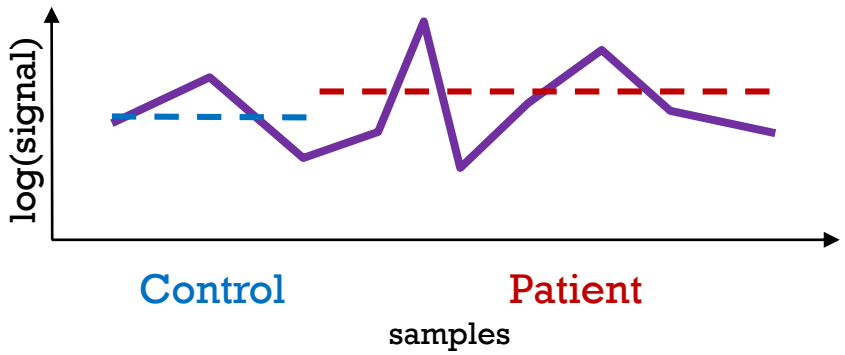
Differential expression analysis: expression signal comparison between groups

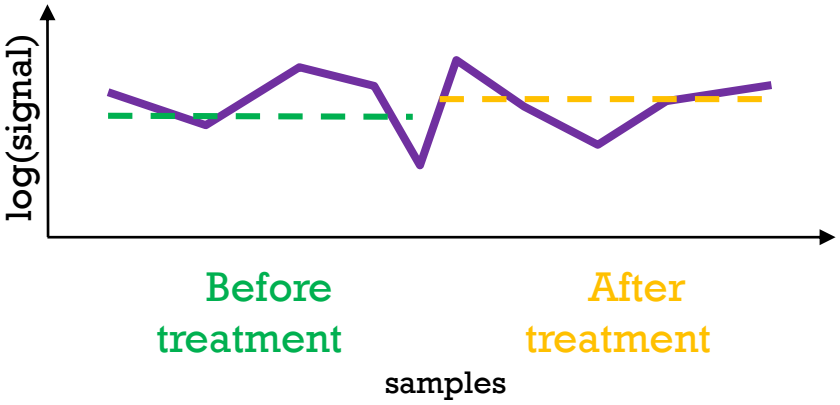
	S_1	S_2	S_3	…	S_{m-1}	S_m
Gene/Protein	X_{11}	X_{12}	X_{13}	…	$X_{1(m-1)}$	X_{1m}
Factor/Group	F_1	F_2	F_3	…	F_{m-1}	F_m

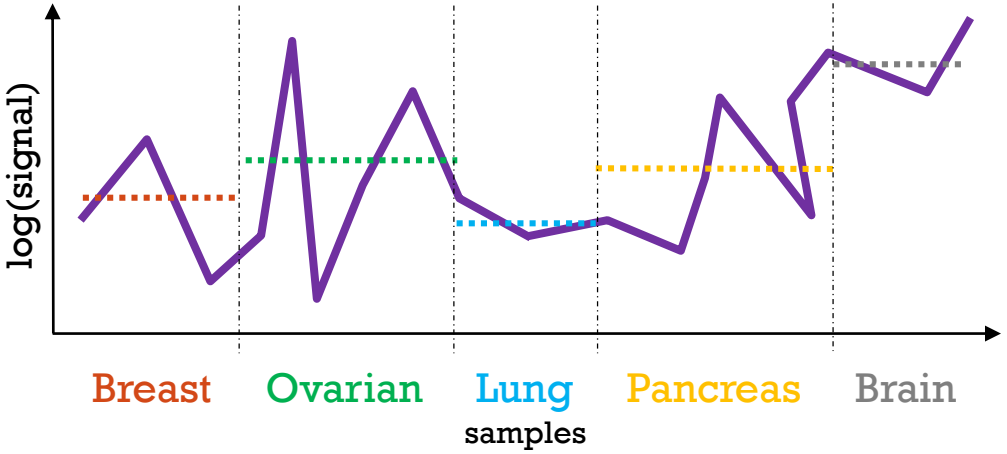


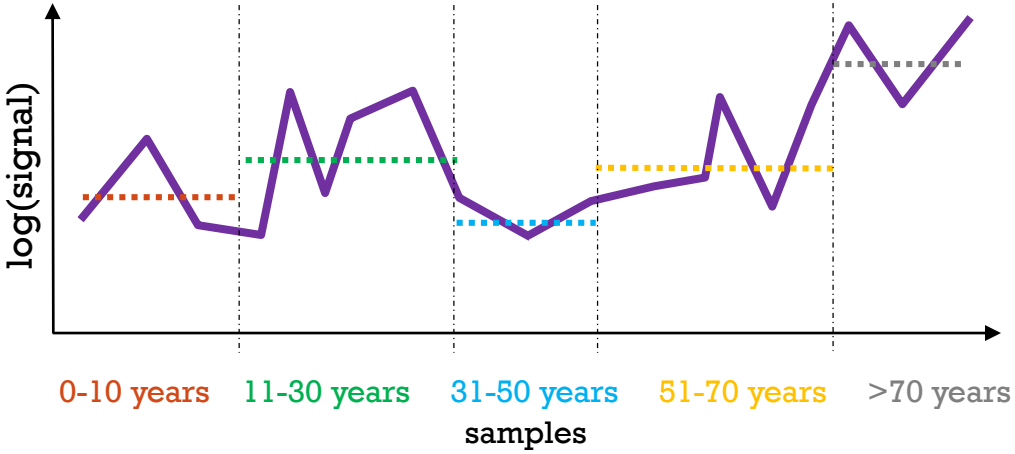
DIFFERENTIAL EXPRESSION

Some scenarios:







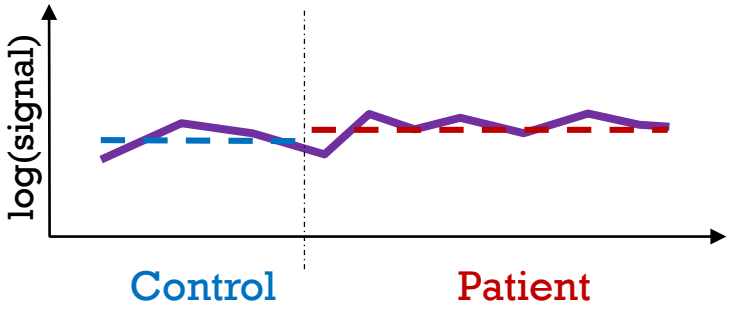


DIFFERENTIAL EXPRESSION

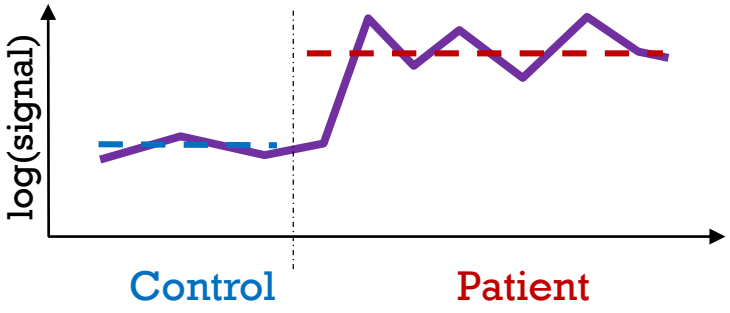
In statistics it is a hypothesis contrast:

H_0: The gene does not change (not DE)

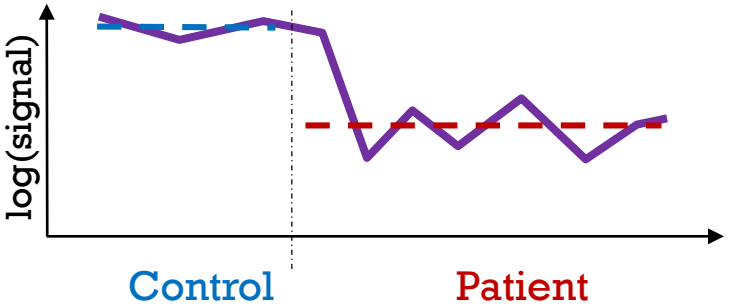
H_1: The gene changes (there are significant differences between groups)



signal(C) ≈ signal(P)
no significant gen



signal(C) < signal(P)
over-expressed gen



signal(C) > signal(P)
under-expressed gen

DIFFERENTIAL EXPRESSION

Objective:

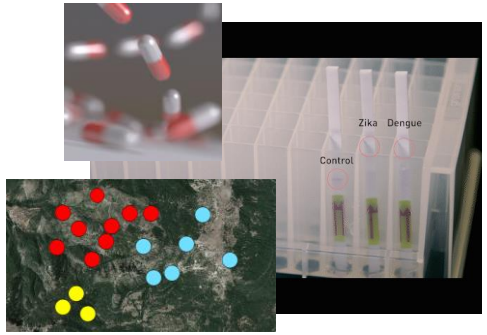
- Find genes that change between groups.

Over-expressed genes → the disease activates it

Under-expressed genes → the disease represses it

* Examples:

- Use this information to activate or repress genes with drugs to fight the disease.
- Find genetic signatures, groups of genes that represent a disease, a type of cancer, a group of individuals…



ACKNOWLEDGMENTS

Group CIC-IBMCC-USAL
Javier De Las Rivas
Diego Alonso
Alberto Berral
Fernando Bueno
Santiago Bueno
Óscar González
Elena Sánchez
José Manuel Sánchez

CIC
CENTRO DE INVESTIGACIÓN DEL CÁNCER

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL