Bioinformatics tool: Standalone interface – R package (Bioconductor)

« ***BiomiX***, *a user-friendly bioinformatic tool for automatized multiomics data analysis and integration* »

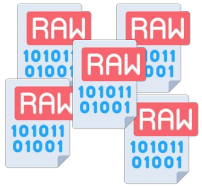
LBAI: [Lymphocytes B, Autoimmunité et Immunothérapies - UMR 1227]

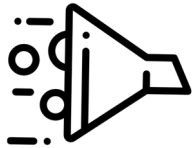
Authors: Cristian Iperi ***Supervisors***: *Álvaro Fernández-Ochoa, Anne Bordon, Christophe Jamin*

Why BiomiX?

The usage of high-throughput technology *in health and biological sciences boosted the amount of information obtainable from samples, ensuring highly robust disease diagnosis and consistent research approaches. The increased dependency on these technologies* ***revealed how data analysis represents the bottleneck step both in time and in skilled bioinformatics users***.

The ***BiomiX*** *offers an efficient and fast pipeline to* ***analyze -omics data individually and integrate multi-omics data*** *from the same patients* ***within the same tool***.



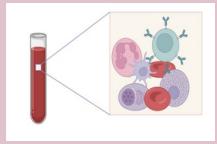


The Input

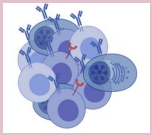
Pipeline on PRECISESADS consortium datasets

Raw matrices + Metadata tables

Transcriptomic data*

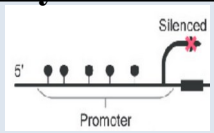


OR



Whole blood RNA-seq ↓ B lymphocytes RNA-seq

Methylomics data*



Whole blood methylomics ↓

Metabolomic data*

OR

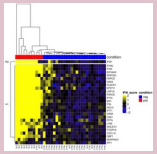
Urine metabolomic ↓ Serum metabolomic

***Suitable also for samples from other sources**

The single -omics Pipeline

Deseq2[2]

DGE analysis



Population and subpopulations analysis based on the expression of a **panel of genes** insert by the users. E.g IFN-α genes for SLE patients

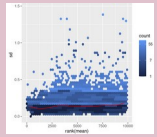
SLE vs CTRL

SLE IFN-α positive vs CTRL

SLE IFN-α negative vs CTRL

Gene variance computation

Vst normalization



Normalized data

Results (p.adj, Log2FC, etc)

Expression matrix in gct format

List of genes up/down regulated

Pathways - biological process discovery

GSEA Gene Set Enrichment Analysis

Enrichr

ChAMP[5]

Variance computation

Data matrix

Beta values matrix

Annotated Results (p.adj, Log2FC, etc)

In-house metabolomics pipeline

Logaritmic normalization

Data matrix

Normalized data

Statistic analysis

Results (p.adj, Log2FC, etc)

Peaks annotation

CEU Mass mediator annotation using more than 7 databases (Kegg, HMDB, LipidMaps, Metlin, NP Atlas, KNApSAcK, MINE). API connection in R using **Cmmr**

Annotation filtering

Human metabolome database predicted, expected and detected metabolites in urine and serum were used to filter the results from the peaks annotation

Annotated Results (p.adj, Log2FC, etc)

Interpretation

Suitable also for samples from other sources

The integration Pipeline

Transcriptomic data*

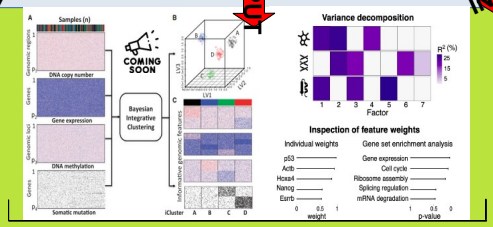
Methylomics data*

Metabolomic data*

INPUT

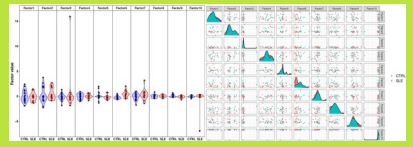
INPUT

INPUT

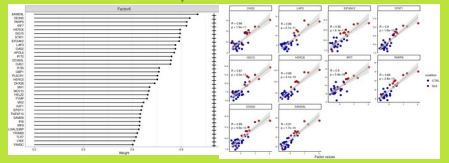


IclusterPlus **MOFA**

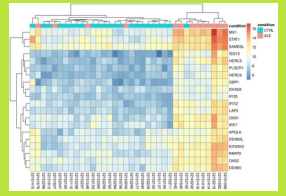
Multiomics data integration



Identification discriminant factors



Identification features contributing to discriminant factors



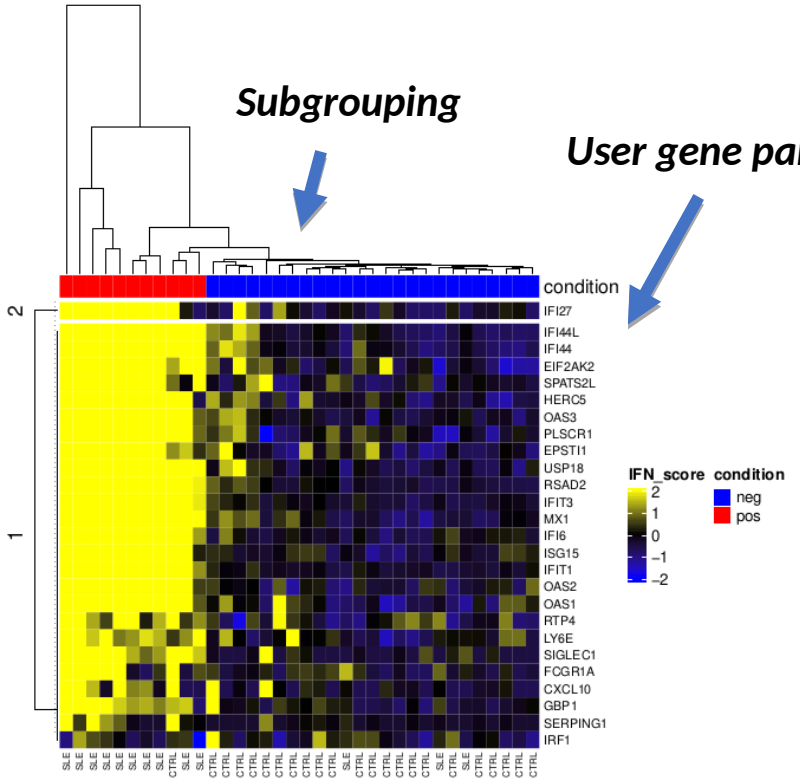
Clustering and heatmap based on discriminant factors

Transcriptomics analysis in details

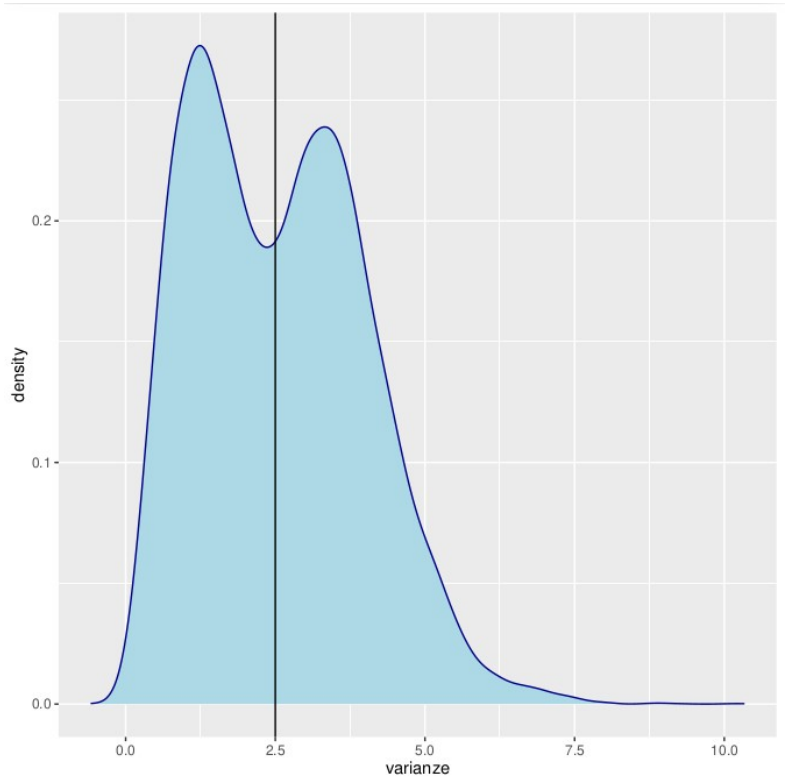
The differential gene expression analysis is made using Deseq2*. The threshold is set to provide results with Log2FC > |0.5| and adjusted p.value < 0.05.*

The ***panel of genes*** *allows* ***to study subpopulations together with the whole population*** *(e.i Interferon-alpha positive/negative). For each comparison the output includes the results file, the data matrix in .gct format for the* ***GSEA analysis*** *and the list of the genes up/downregulated for the* ***EnrichR*** *analysis. Also, a* ***volcano plot*** *is drawn for each condition*

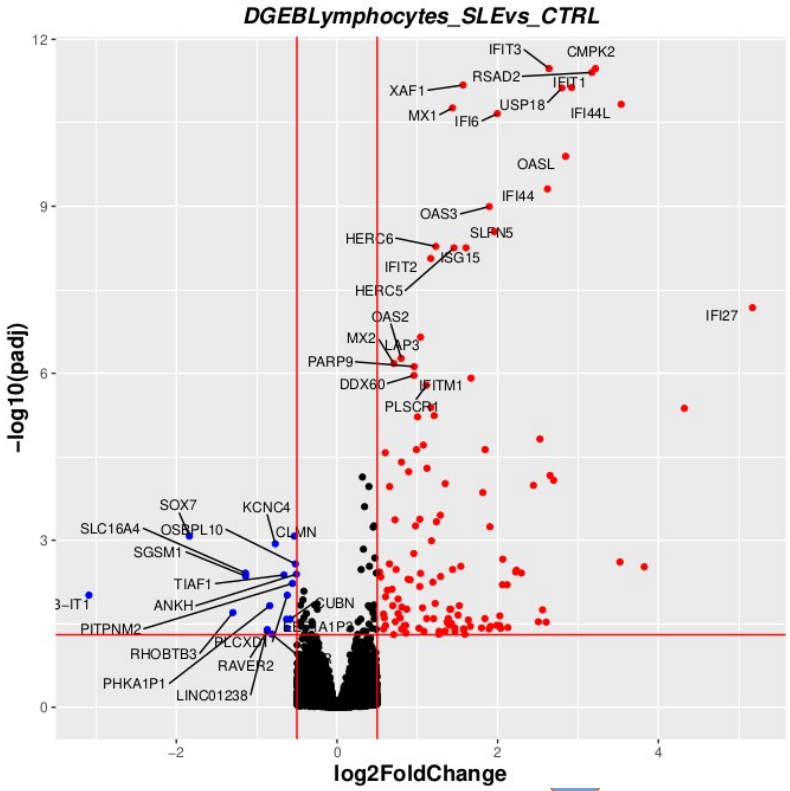
The subgrouping plot using the gene panel is provided, also with threshold parameters to increase and improve the subgrouping sensibility. It ***is possible to use an external marker to validate*** *the belonging of the* ***subgroups.*** *A* ***plot*** *resuming the* ***genes variances*** *is also produced.*



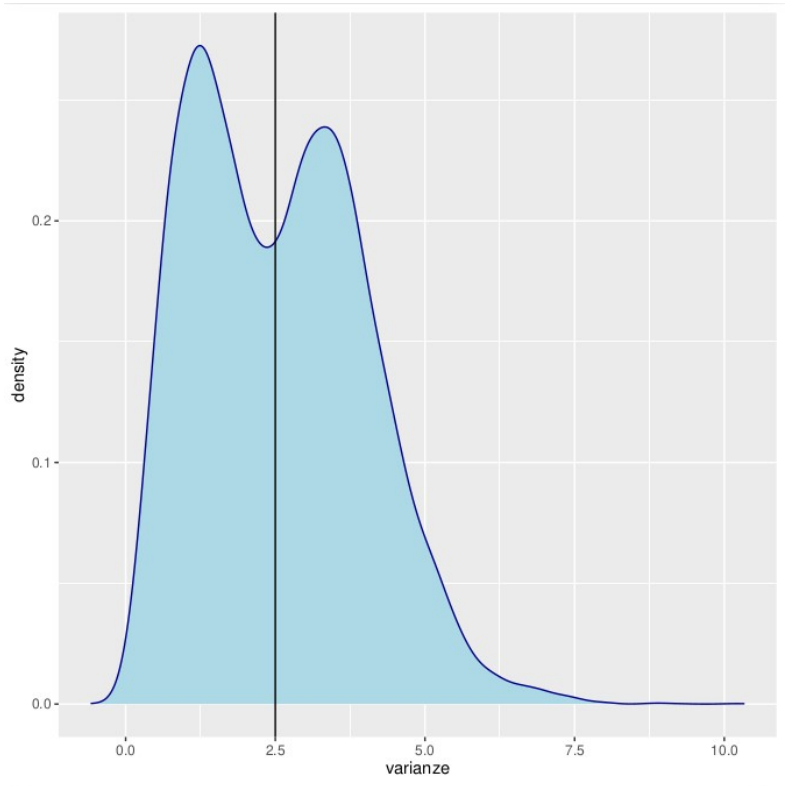
Subgrouping patients by gene panel



Gene variance distribution



Volcano plot differential gene expression (ggplot)



Gene variance distribution

Metabolomics analysis in details

The peaks signal analysis is made using an in-house script calculating Lo2FC and adjusted p.values from Mann-Whitney corrected with FDR. *. The threshold is set to provide results with Log2FC > |0.5| and adjusted p.value < 0.05.*

The script seeks a match between the peaks annotations and the 7 databases linked to ***CEU MASS MEDIATOR*** *(Kegg, HMDB, LipidMaps, Metlin, NP Atlas, KNApSAcK, MINE) and retrieves the candidate metabolites. The* ***candidate metabolites in urine and serum metabolomics can be filtered*** *using the metabolites previously predicted or detected in the scientific community,* ***to reduce the number of candidates****.*

The output file contains adducts, names, mass and links for the 7 databases for each metabolite.

Name	Formula	Adduct	Molecular.Weight	PPM.Error	Ionization.Score	Final.Score	Kegg	Kegg_URI	HMDB_URI	LipidMaps	LipidMaps_URI
Hydroxychloroquine	C18H26ClN3O	M+H	335.176440176	1	-2	-2	C07043	http://www.genome.jp/d	http://www.hmdb.ca/metabolit	NA	NA
OR-1896	C13H15N3O2	M+H	245.116426739	1	-2	-2	NA	NA	http://www.hmdb.ca/metabolit	NA	NA
PC(14:0/22:5(4Z,7Z,10Z,13Z,16Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01012130	http://www.lipidmaps.org/d
PC(14:0/22:5(7Z,10Z,13Z,16Z,19Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01012131	http://www.lipidmaps.org/d
PC(14:1(9Z)/22:4(7Z,10Z,13Z,16Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01011407	http://www.lipidmaps.org/d
PC(16:0/20:5(5Z,8Z,11Z,14Z,17Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01010633	http://www.lipidmaps.org/d
PC(16:1(9Z)/20:4(5Z,8Z,11Z,14Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01010695	http://www.lipidmaps.org/d
PC(16:1(9Z)/20:4(8Z,11Z,14Z,17Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01012139	http://www.lipidmaps.org/d
PC(18:1(11Z)/18:4(6Z,9Z,12Z,15Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01012152	http://www.lipidmaps.org/d
PC(18:1(9Z)/18:4(6Z,9Z,12Z,15Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01011604	http://www.lipidmaps.org/d
PC(18:2(9Z,12Z)/18:3(6Z,9Z,12Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01011625	http://www.lipidmaps.org/d
PC(18:2(9Z,12Z)/18:3(9Z,12Z,15Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01011626	http://www.lipidmaps.org/d
PC(18:3(6Z,9Z,12Z)/18:2(9Z,12Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01011653	http://www.lipidmaps.org/d
PC(18:3(9Z,12Z,15Z)/18:2(9Z,12Z))	C44H78NO8P	M+Na	779.546504989	8	1	1.489675	NA	NA	http://www.hmdb.ca/metabolit	LMGP01011683	http://www.lipidmaps.org/d

Example partial metabolomics results table

Methylomics analysis in details

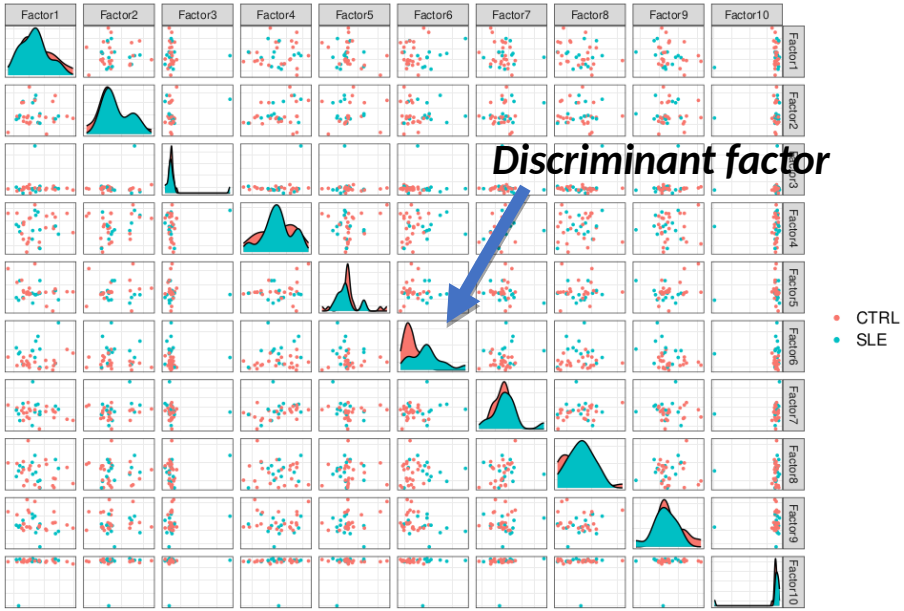
The evaluation of the different methylations in the CpG island was made using ChAMP. And so the annotation and the linkage with the genes. For further information, we suggest you take a look at the article and the package.

https://www.bioconductor.org/packages/devel/bioc/vignettes/ChAMP/inst/doc/ChAMP.html

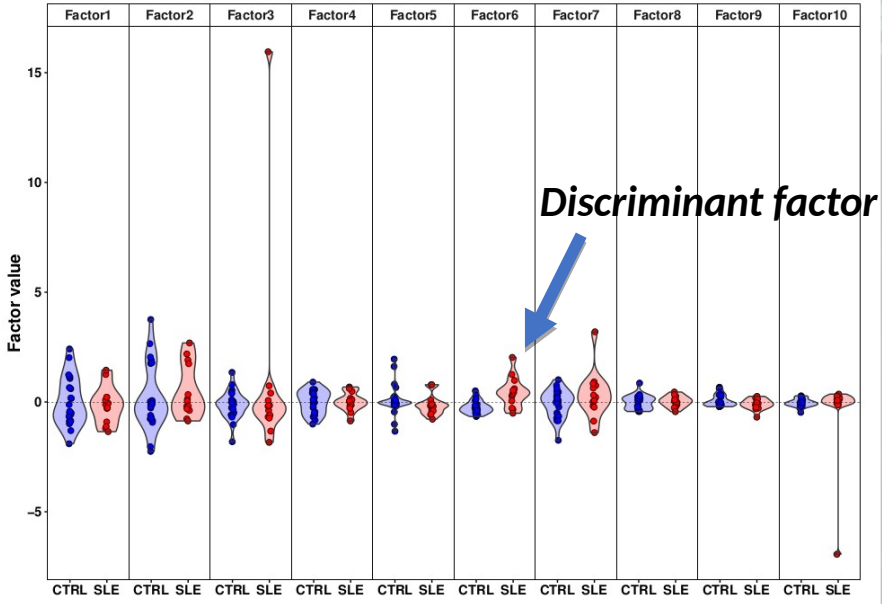
MOFA analysis in details

The usage of ***MOFA is based on the identification of factors to distinguish the two conditions/populations*** *that we are interested in. The tool allows selecting the number of factors calculated and the ones on which we want to focus further analysis (****discriminating factor****).*

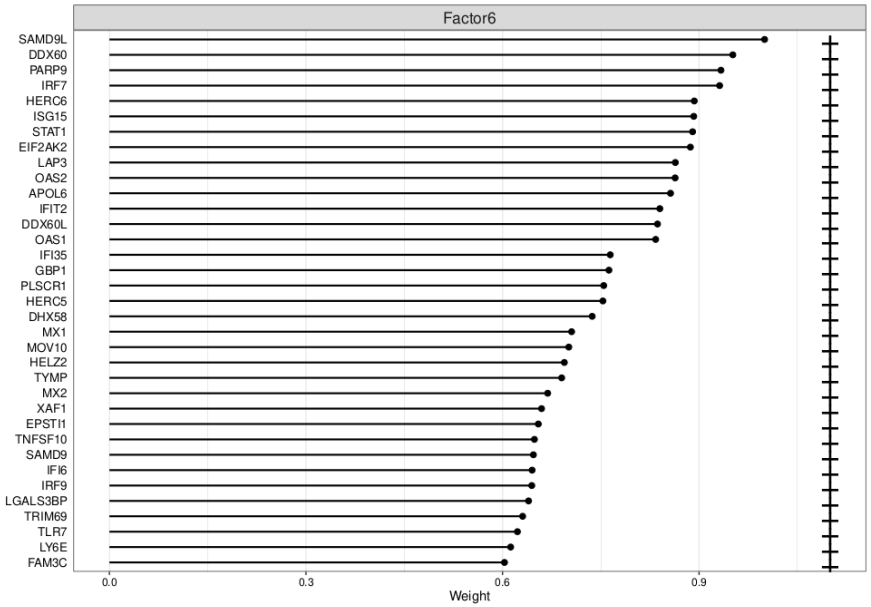
MOFA allows ***testing the clustering efficiency*** *of our conditions* ***in the –Omics contributing to the chosen discriminating factor****, by heatmaps. It also shows the top features (genes/peaks/CpG island) with higher weight in the selected factor.*

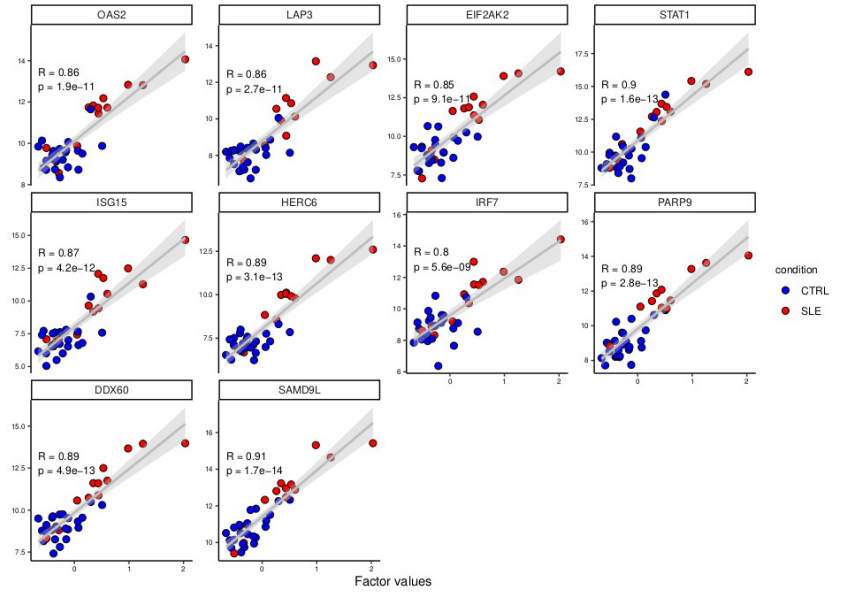


Gggally plot to select the discriminant factor

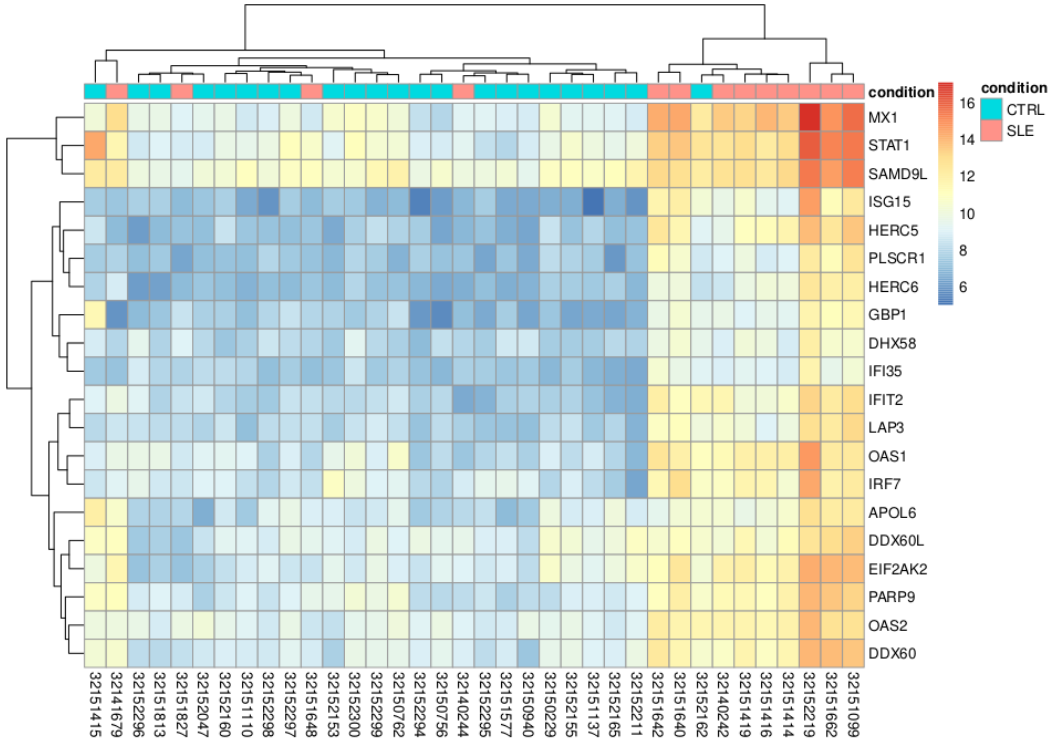


Violin plot to select the discriminant factor





B cell top genes contributing to discriminant factor and their scatter plot



Whole blood heatmap and clustering of top genes contributing to discriminant factor

Further information and usage?

BiomiX** is **writen in R** language, and the **interface in python.

If you are interested** in using it please **write your e-mail and name into the google document here.

https://docs.google.com/spreadsheets/d/10V0QF5IeqdT5fhG_nhh-NLuyGzWELbFNLWzP4a3YwTc/edit?usp=sharing

***If you want to use the beta-version,** please contact me:*

cristian.iperi@univ-brest.fr/cristian.iperi@gmail.com