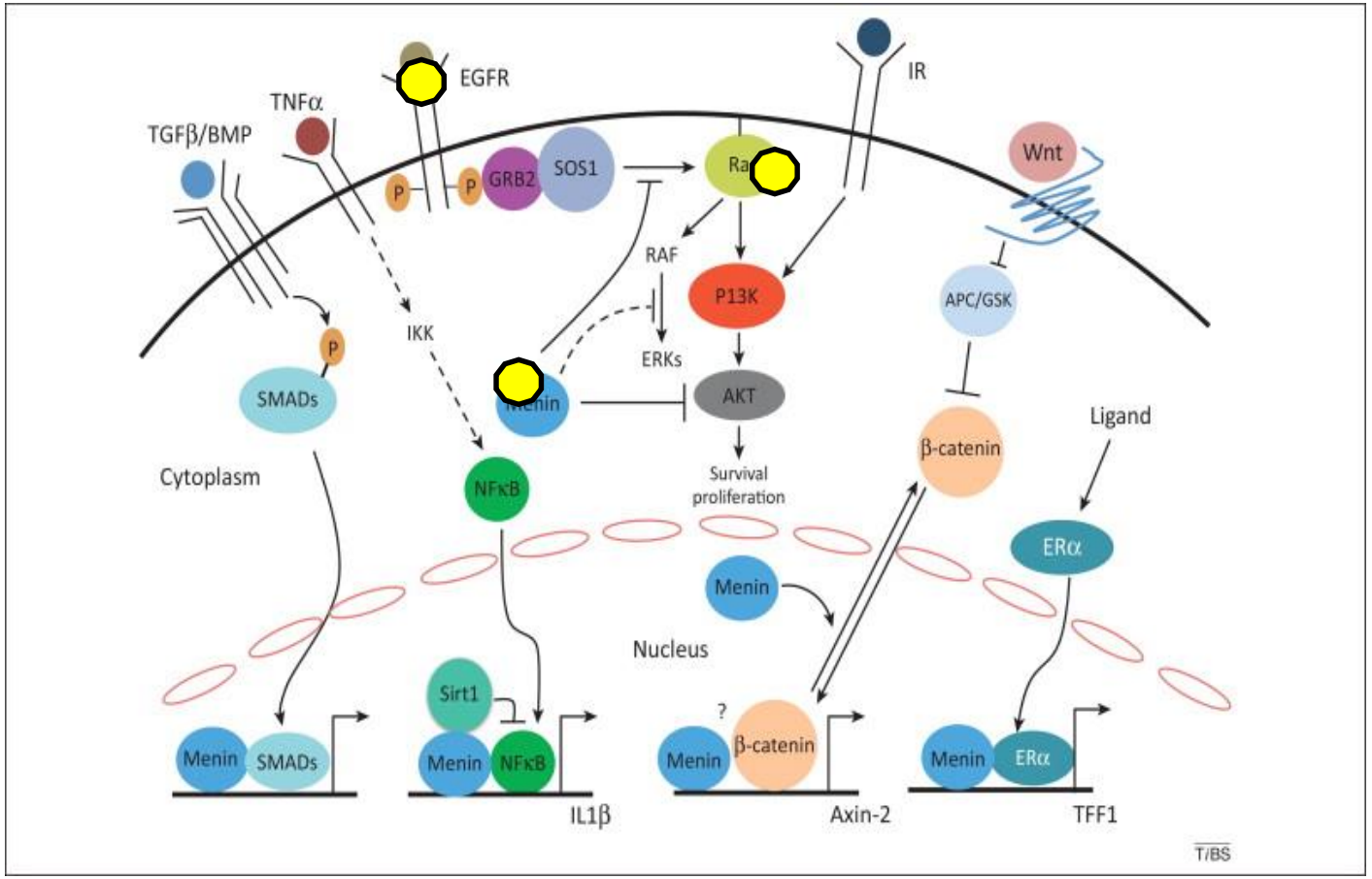
From genes to pathways:

pathway quantification with ROMA

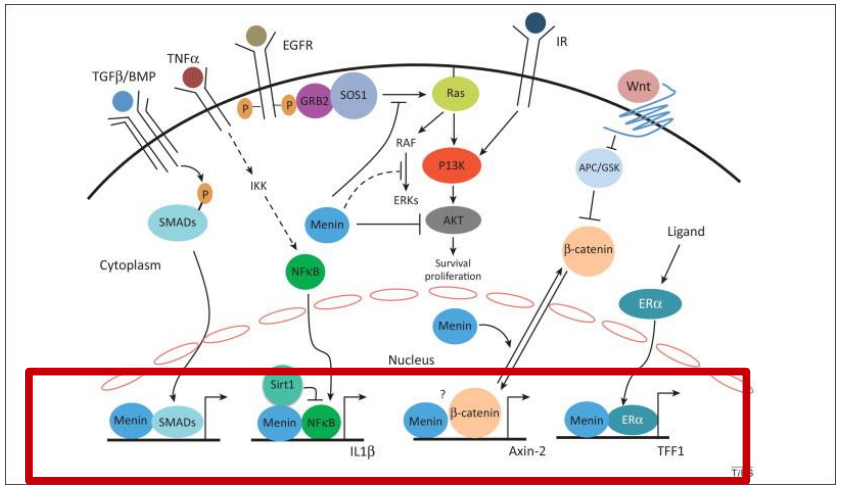
A Montagud, L Albergante, U Czerwinska, A Zinovyev, L Martignetti
U900 Computational Systems Biology of Cancer team
Institut Curie

In cancer the same biological process can be affected by damages in different individual genes

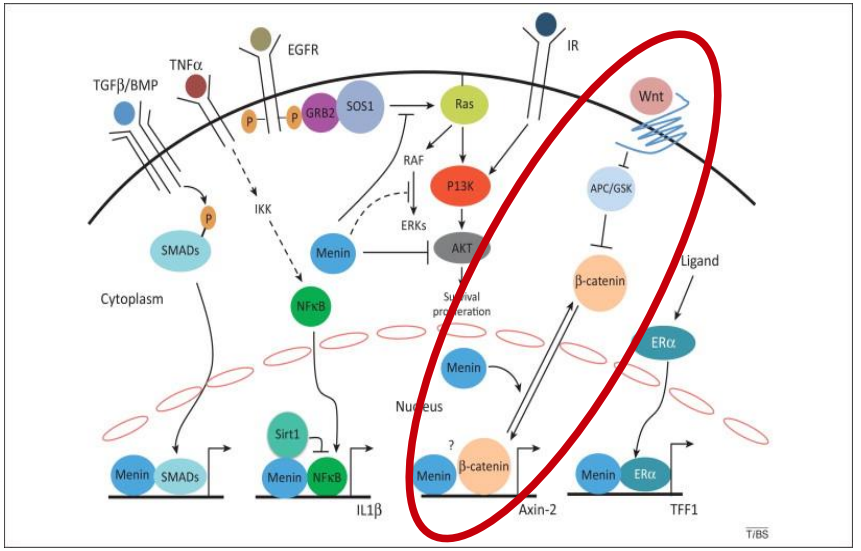


alterations

Reasoning in terms of active/inactive gene-sets rather than single differentially expressed genes



Gene set = target genes co-regulated by the same TFs



Gene set = genes involved in a common signalling pathway

Pathway-level analysis :

- make use of existing knowledge (e.g. public database, literature, etc.)

- try to "separate scales", identify and retain coarse-grained variables that are essential for the problem

Quantification of gene-set activity

→ Single biomarker gene expression as a proxy of the whole gene-set

→ Mean/Median expression of the genes in the set

Some drawbacks :

- Different genes do not contribute in the same way/strength to the activity of the gene-set

- Some genes can correlate negatively with the activity of the gene-set

Alternative :

Gene set activity as a linear combination of individual gene expression

$$A_j = \sum \alpha_i \, x_{ij}$$

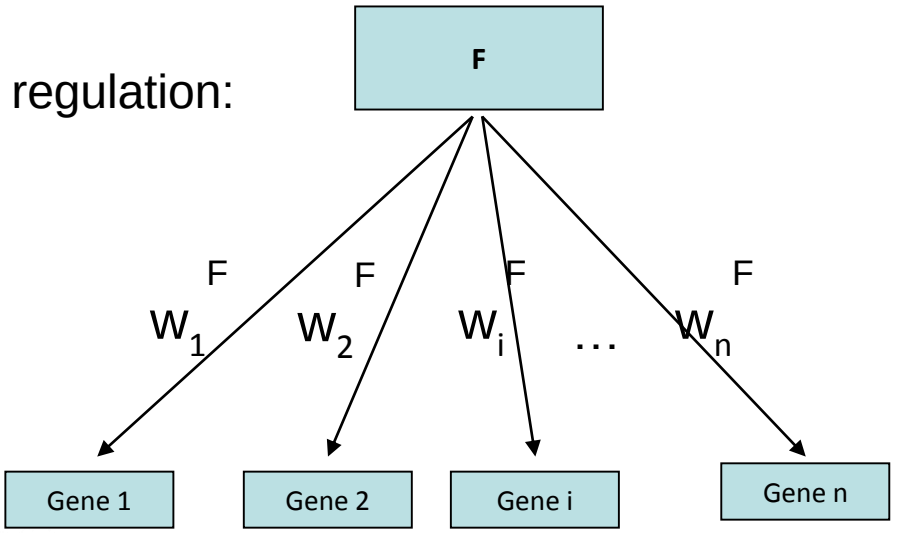
Fan J et al, Nature Methods 2016

Tomfohr et al, BMC Bioinformatics 2005

Quantification of gene-set activity by PCA

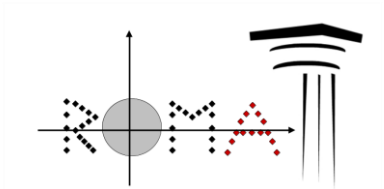
The uni-factor linear model of gene expression regulation:

$$x(gene_i, S_j) \sim w_i^{(F)} A_j^{(F)}$$



frontiers in Genetics

METHODS
published: 19 February 2016
doi: 10.3389/fgene.2016.00018



ROMA: Representation and Quantification of Module Activity from Target Expression Data

Loredana Martignetti[1,2,3,4], *Laurence Calzone*[1,2,3,4], *Eric Bonnet*[1,2,3,4], *Emmanuel Barillot*[1,2,3,4] *and Andrei Zinovyev*[1,2,3,4]*

institutCurie

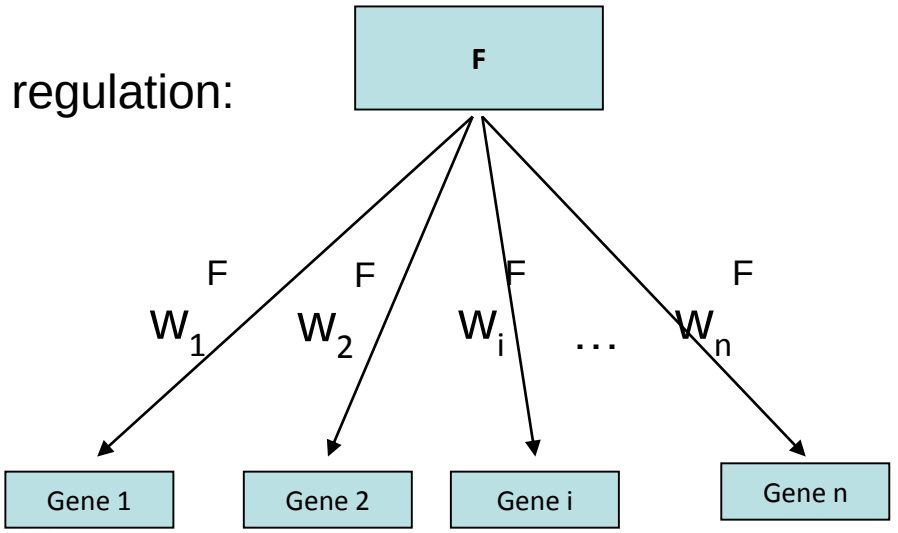
Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Quantification of gene-set activity by PCA

The uni-factor linear model of gene expression regulation:

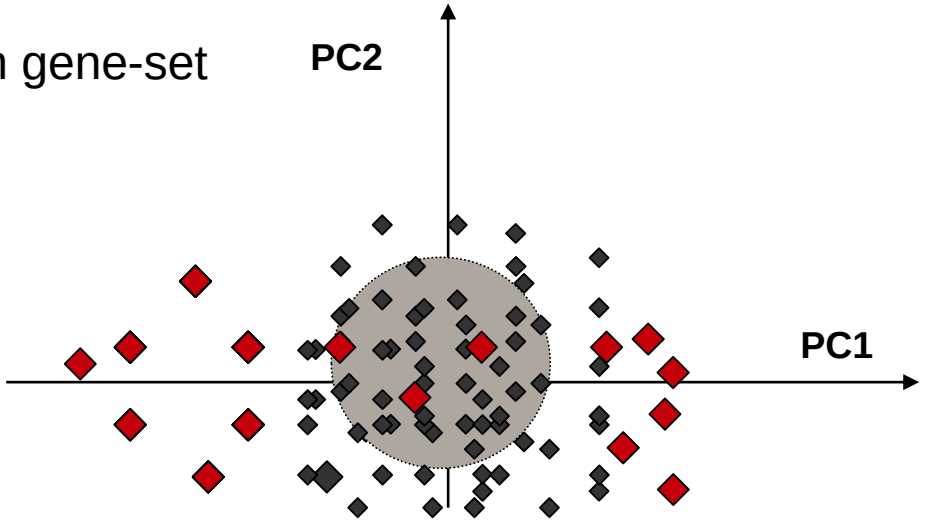


$$x(gene_i, S_j) \sim w_i^{(F)} A_j^{(F)}$$

$$X = W\ D\ A$$

$$A_j^{(F)} \sim \lambda^{-1} \sum_i w_i^{(F)}\, x_{ij}$$

◆ Gene from gene-set of interest



The values $w_i^{(F)}$ and $A_j^{(F)}$ are obtained by the first metagene **PC1 of the gene set** and by the level of this metagene in each sample

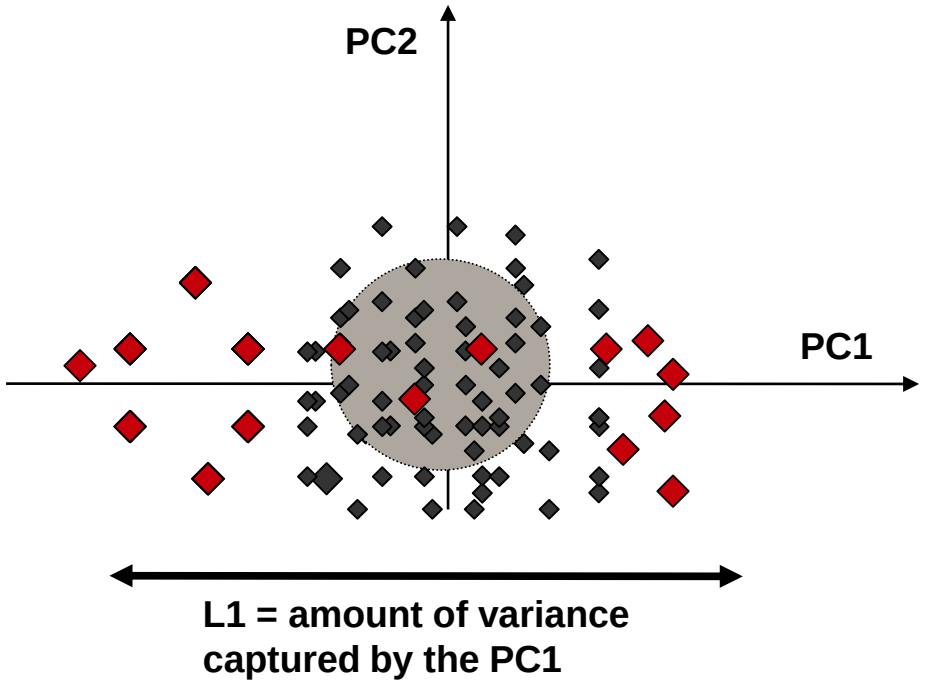
institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

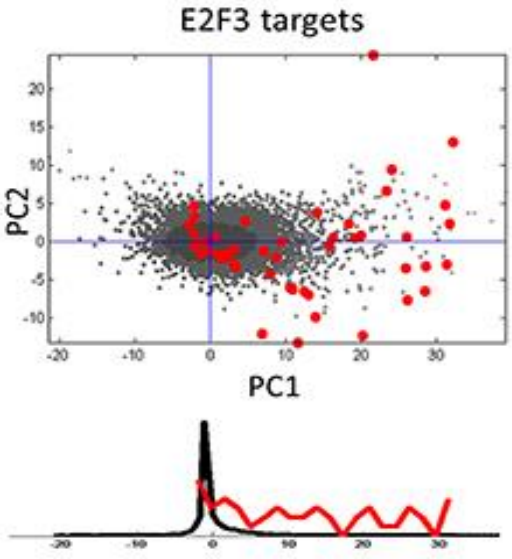
PSL RESEARCH UNIVERSITY

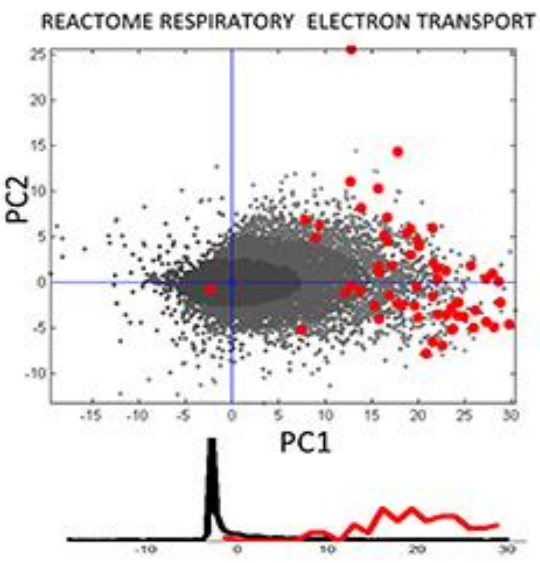
Identification of significantly active/inactive gene-sets by PC1



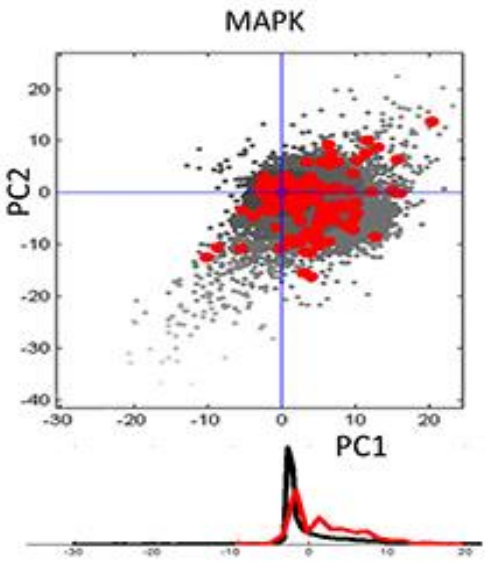
Testing if the PC1 variance L1 of a gene-set significantly exceeds the genome-wide background expectation = overdispersion

Overdispersed gene sets, p-value<0.01





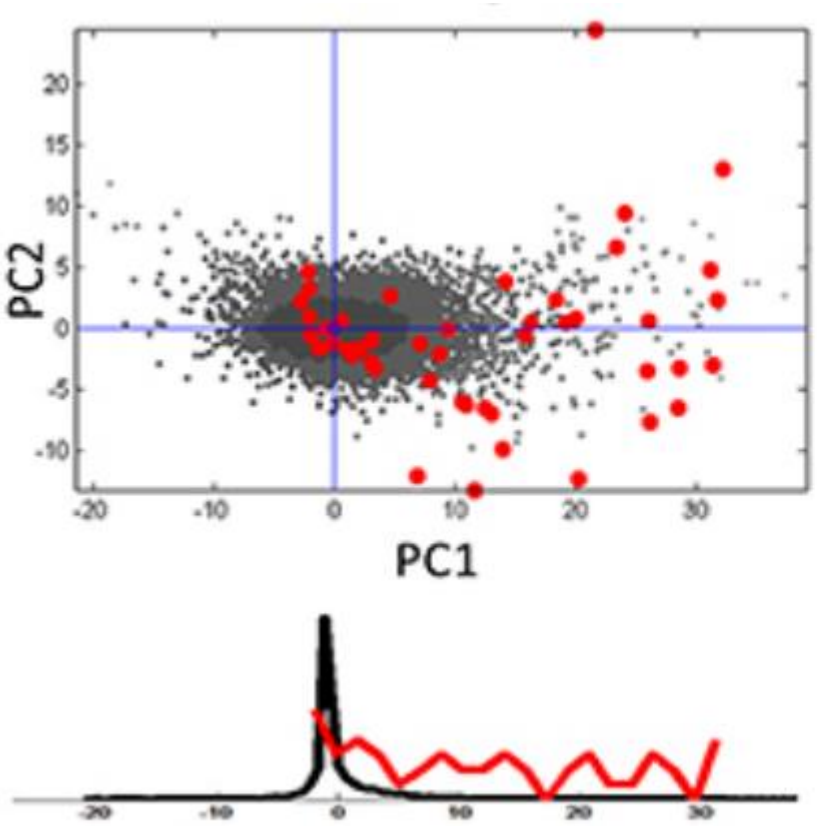
Non-overdispersed



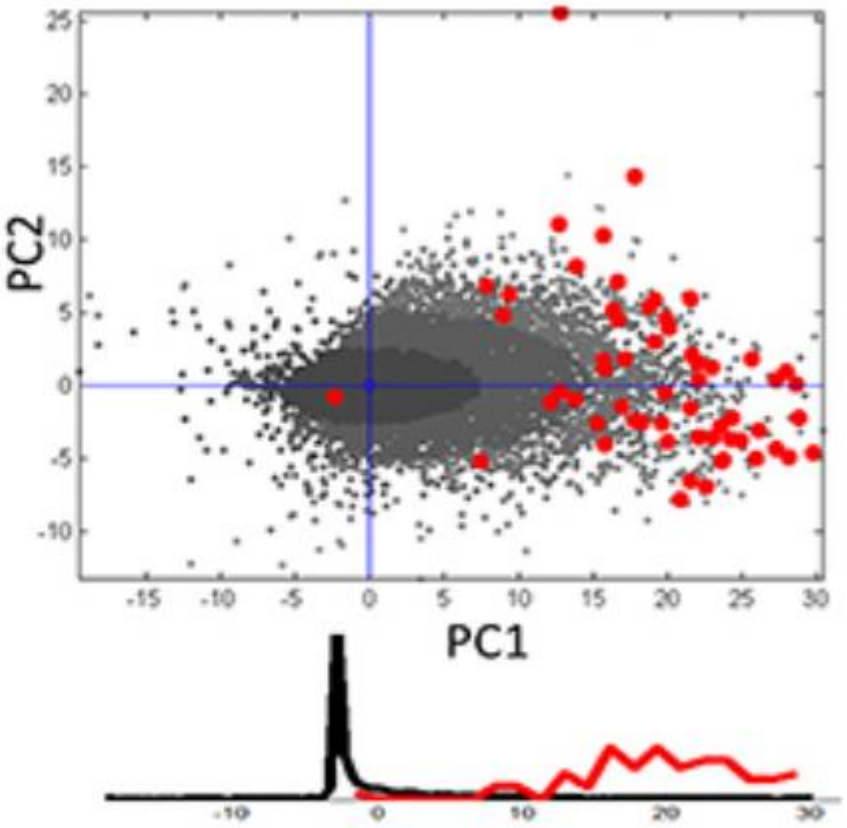
ROMA features: computing PCA wih fixed center

Two possible configurations of the target genes :

1. Only some genes of the modules show overdispersion compared to the background



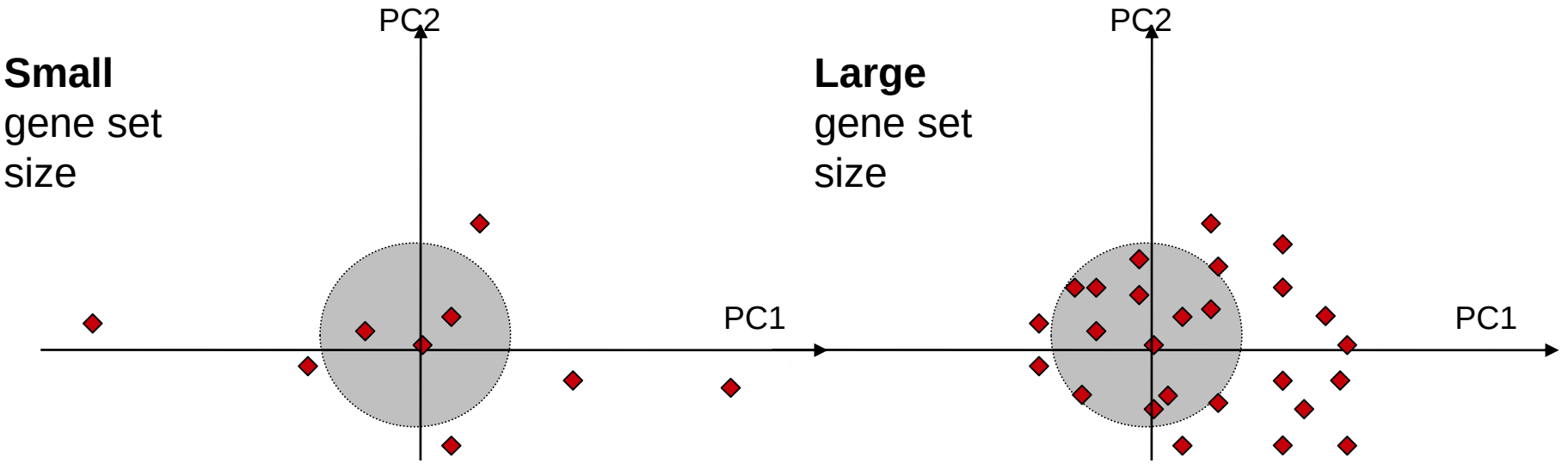
2. All genes of the modules are shifted compared to the background genes



The two configurations can be detected in ROMA using **PCA with fixed center**

ROMA features:
assessing the statistical significance of gene-set overdispersion

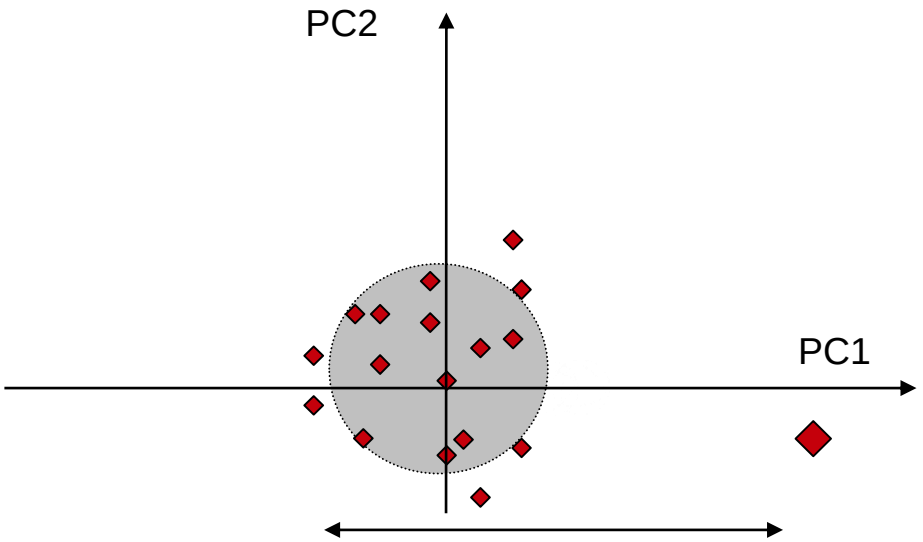
L1 and L1/L2 strongly depend on the size of the gene set



Statistically significance of L1 and L1/L2 is assessed by estimating the null distribution of L1 and L1/L2 from **random set of genes having representative sizes**

ROMA features: computing robust PCA

Outlier genes abnormally affecting PC1 are identified by "leave one out" procedure and removed from the gene-set



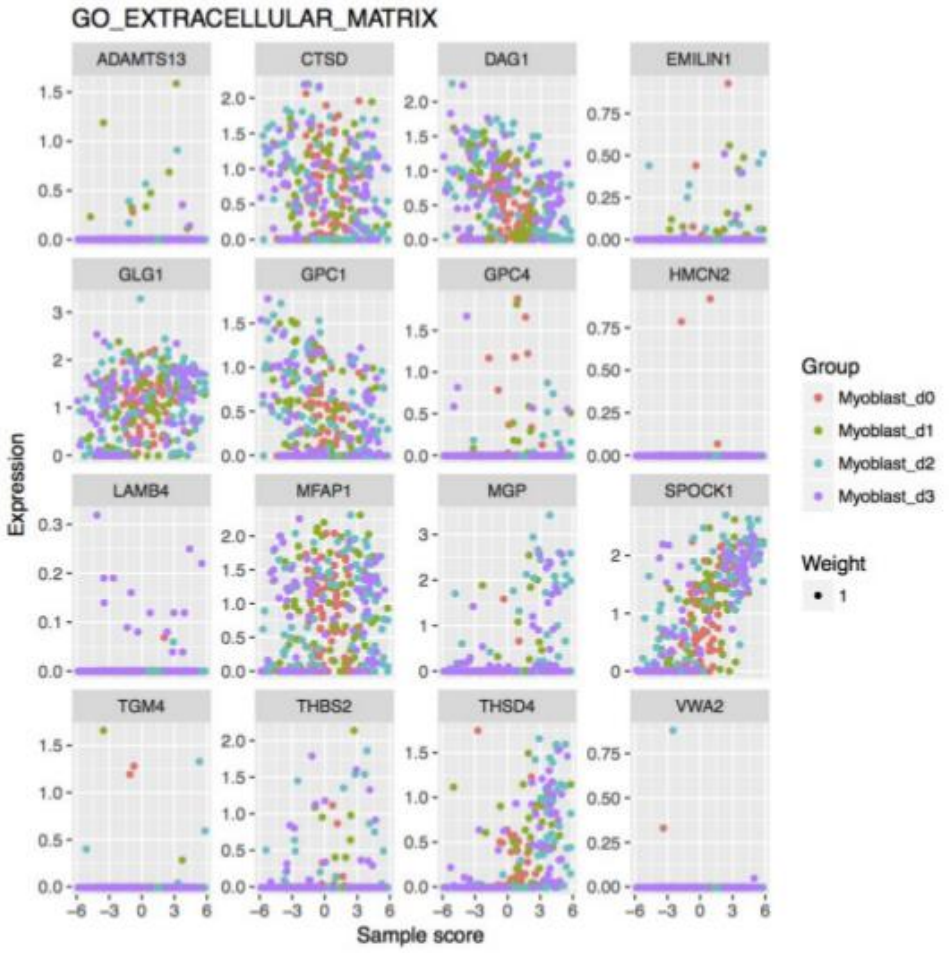
L1 estimate is affected by one single gene

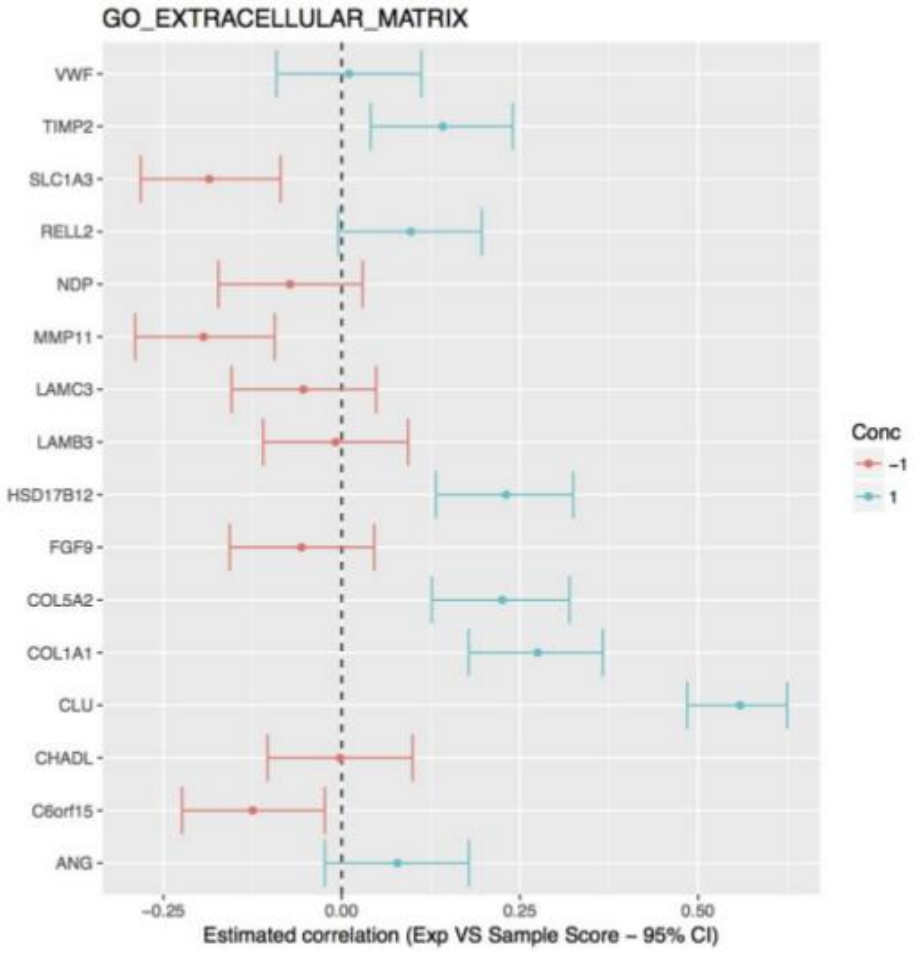
In ROMA outlier genes are identified by leave-one-out procedure:

→ computing L1 *n* times (*n* = gene set size) removing at each time one gene in the gene set

→ outliers are identified as those genes that *dramatically* increase L1

ROMA features: orienting PCA





In ROMA PCA is oriented in such a way that gene projections are positively correlated with gene expression levels for most genes

ROMA features:

using weighted gene-sets to include a priori biological knowledge

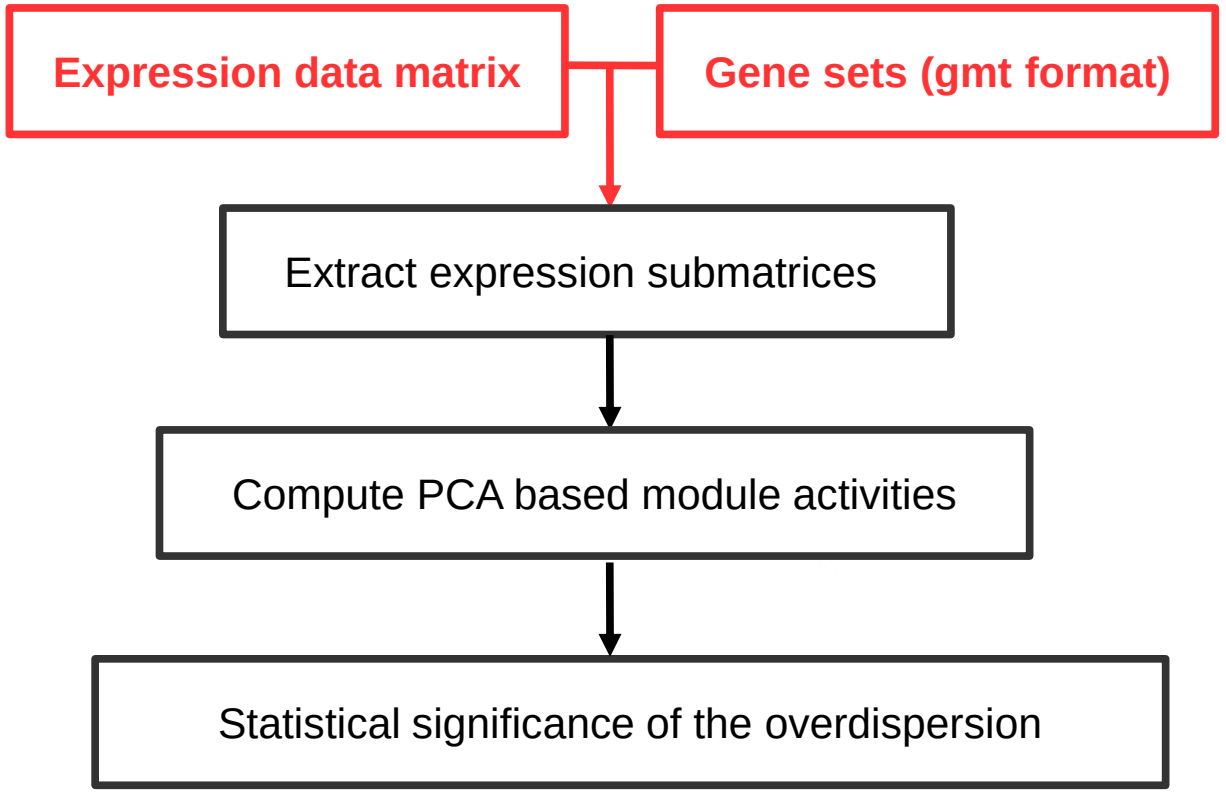
In ROMA, some weights w_g can be assigned by the user (weighted gmt file)

Example:

positive weights for “positively regulated genes” and negative for “inhibited genes”

Bigger weights for user-defined « most contributing » genes of the gene-set

The ROMA algorithm



OUTPUT

Overdispersion estimation L1+ pv
Coordinated expr. estimation L1/L2 + pv

Module activity score in each condition

Gene projections on PC1

https://github.com/sysbio-curie

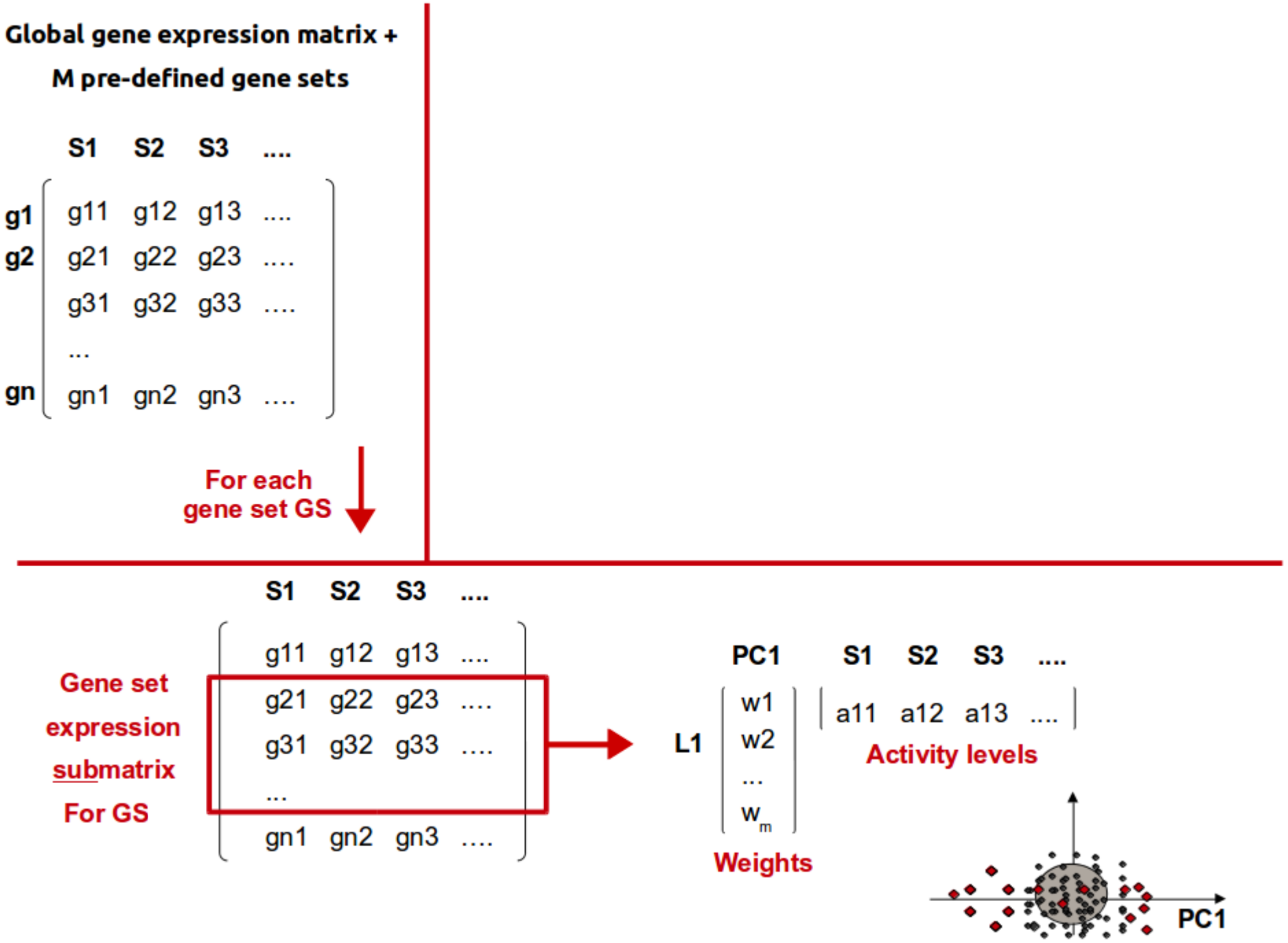
LM et al, Front Genet. 2016

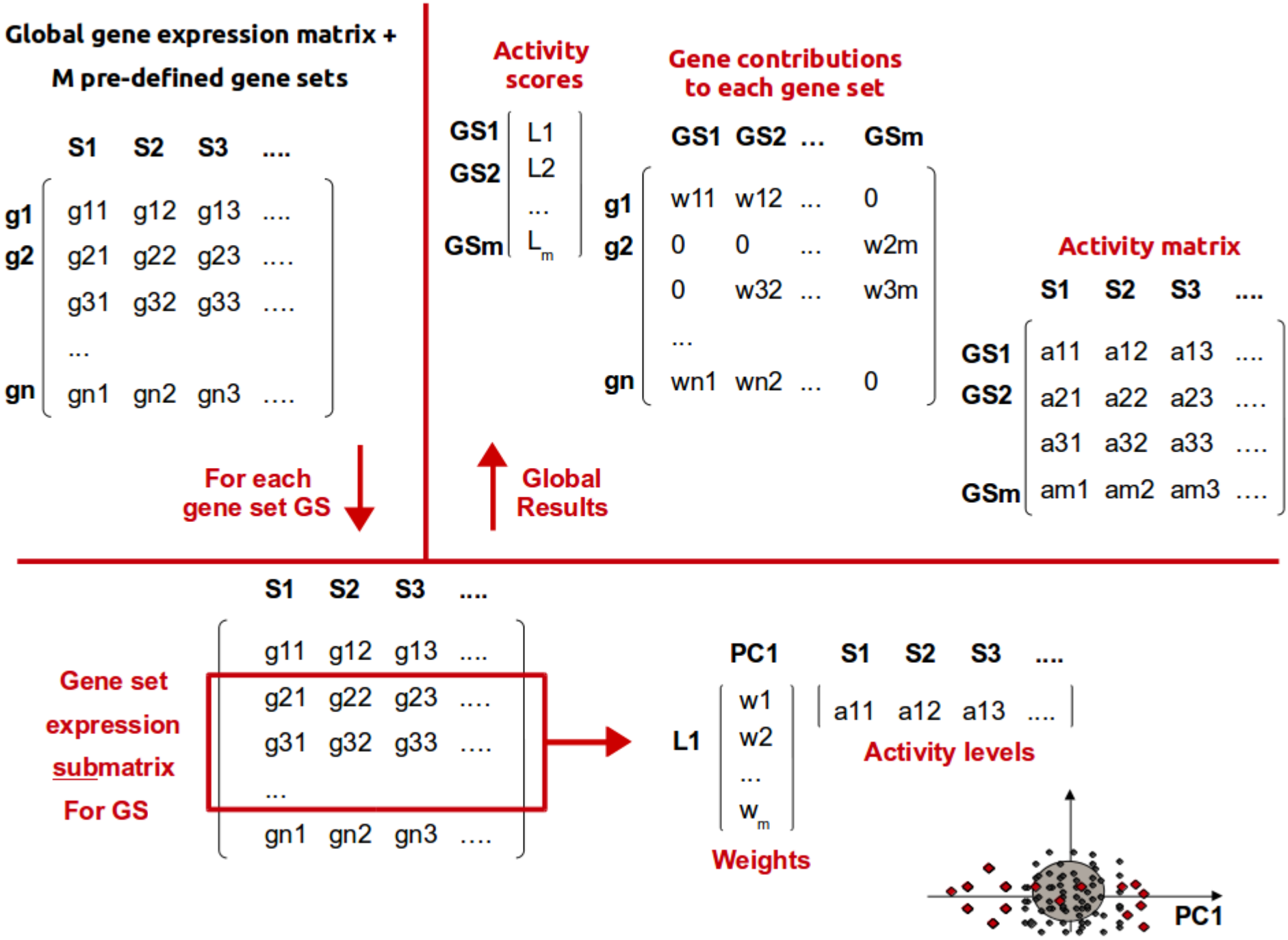
Global gene expression matrix +

M pre-defined gene sets

	S1	S2	S3	....
g1	g11	g12	g13	....
g2	g21	g22	g23	....
	g31	g32	g33	....
	...			
gn	gn1	gn2	gn3	....

For each gene set GS


Global gene expression matrix +
M pre-defined gene sets
S1 S2 S3 ....
g1 g11 g12 g13 ....
g2 g21 g22 g23 ....
g31 g32 g33 ....
...
gn gn1 gn2 gn3 ....
For each
gene set GS
S1 S2 S3 ....
Gene set
expression
submatrix
For GS
g11 g12 g13 ....
g21 g22 g23 ....
g31 g32 g33 ....
...
gn1 gn2 gn3 ....
PC1 S1 S2 S3 ....
L1 w1 w2 ... wm
a11 a12 a13 ....
Activity levels
Weights
PC1


Global gene expression matrix +
M pre-defined gene sets
S1 S2 S3 ....
g1 g11 g12 g13 ....
g2 g21 g22 g23 ....
g31 g32 g33 ....
...
gn gn1 gn2 gn3 ....
For each gene set GS
Activity scores
GS1 L1
GS2 L2
...
GSm Lm
Global Results
Gene contributions to each gene set
GS1 GS2 ... GSm
g1 w11 w12 ... 0
g2 0 0 ... w2m
0 w32 ... w3m
...
gn wn1 wn2 ... 0
Activity matrix
S1 S2 S3 ....
GS1 a11 a12 a13 ....
GS2 a21 a22 a23 ....
a31 a32 a33 ....
GSm am1 am2 am3 ....
Gene set expression submatrix For GS
S1 S2 S3 ....
g11 g12 g13 ....
g21 g22 g23 ....
g31 g32 g33 ....
...
gn1 gn2 gn3 ....
PC1
L1 w1 w2 ... wm
Weights
S1 S2 S3 ....
a11 a12 a13 ....
Activity levels
PC1

How to use ROMA in practice

Java version @ https://github.com/sysbio-curie/Roma

Command line usage:

java -jar roma_v1.0.jar [required options] [other options]

R version @ https://github.com/sysbio-curie/rRoma

R version using shiny dashboard @
https://github.com/sysbio-curie/rRomaDash

Execute rRoma with command line

Load data

Expression matrix file

Sample annotation file

Testing different signatures for a given pathway (ex: wnt)

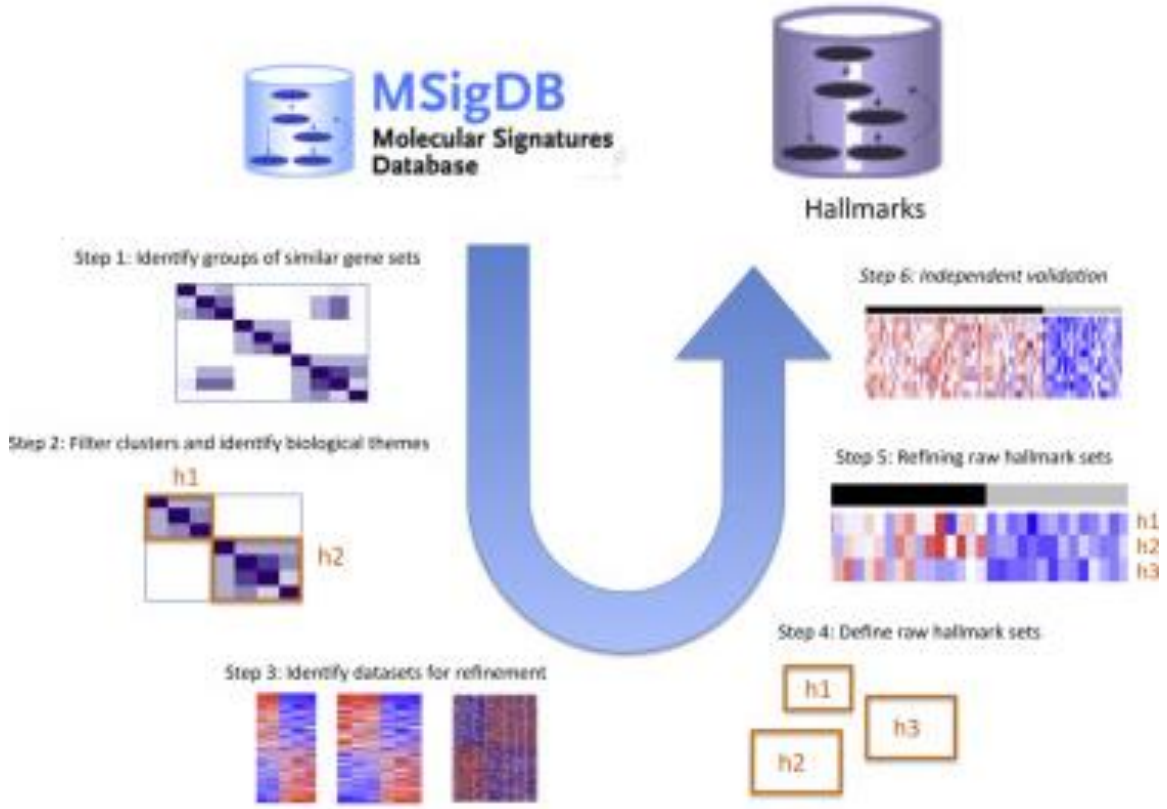
```
wntGMT <- ReadGMTFile("Unsigned_wnt_path.gmt")

Data.wnt <- rRoma.R(ExpressionMatrix = expr,
                    ModuleList = wntGMT,
                    FixedCenter = TRUE,
                    MaxGenes = 1000,
                    PCSignMode="CorrelateAllWeightsByGene",
                    PCAType = "DimensionsAreSamples")
```

Results of different signatures of WNT pathway

	L1	ppv L1	L1/L2	ppv L1/L2
WNT_CANONICAL	0.2160966	0.45	1.571934	0.61
Wnt_CELL_MAP	0.3168016	0.18	2.465684	0.10
WNT_NON_CANONICAL	0.2237947	0.34	1.684310	0.48
WNT_pthw_Metastasis	0.3099135	0.04	2.173662	0.14
wnt_IPA	0.3020718	0.00	2.734551	0.01

Two WNT signatures (WNT_pthw_Metastasis,wnt_IPA) perform better than the others

The Molecular Signatures Database (MSigDB) hallmark gene set collection



Hallmark gene sets represent specific well-defined biological states or processes that display coherent expression

* Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. **The Molecular Signatures Database (MSigDB) hallmark gene set collection**. Cell Syst. 2015 Dec 23;1(6):417-425.

institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Testing a database of signatures (ex: MsigDB Hallmarks)

```
AllHall <- SelectFromMSIGdb("HALLMARK")


Data.hall <- rRoma.R(ExpressionMatrix = expr,
                     ModuleList = AllHall,
                     FixedCenter = TRUE,
                     MaxGenes = 1000,
                     PCSignMode="CorrelateAllWeightsByGene",
                     PCAType = "DimensionsAreSamples")
```

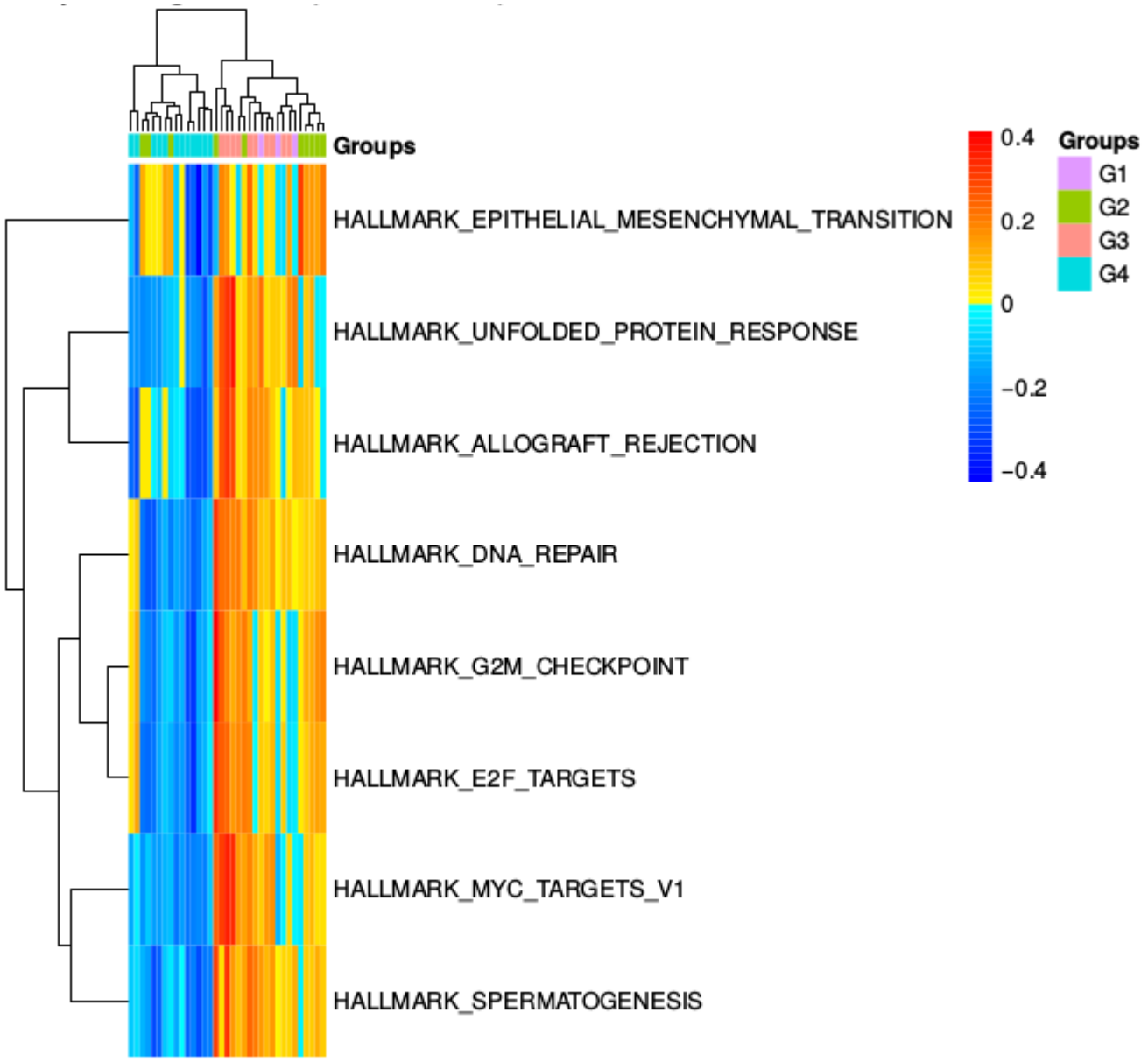
Selecting significantly active/inactive modules

```
AggData.FC <- Plot.Genesets (RomaData = Data.hall,
        Selected = SelectGeneSets (RomaData = Data.hall,
                                   VarThr = 1e-05,
                                   VarMode = "Wil",
                                   VarType = "Over"),
        GenesetMargin = 20,
        SampleMargin = 14,
        cluster_cols = TRUE,
        GroupInfo = Group,
        AggByGroupsFL = c("mean", "sd"),
        HMTite = "Overdispersed genesets (Fixed center)")
```

-> over- or under-underdispersed genesets selected according to VarThr p-value threshold and VarMode = "Wil" or "PPV" for Wilcoxon or permutation test

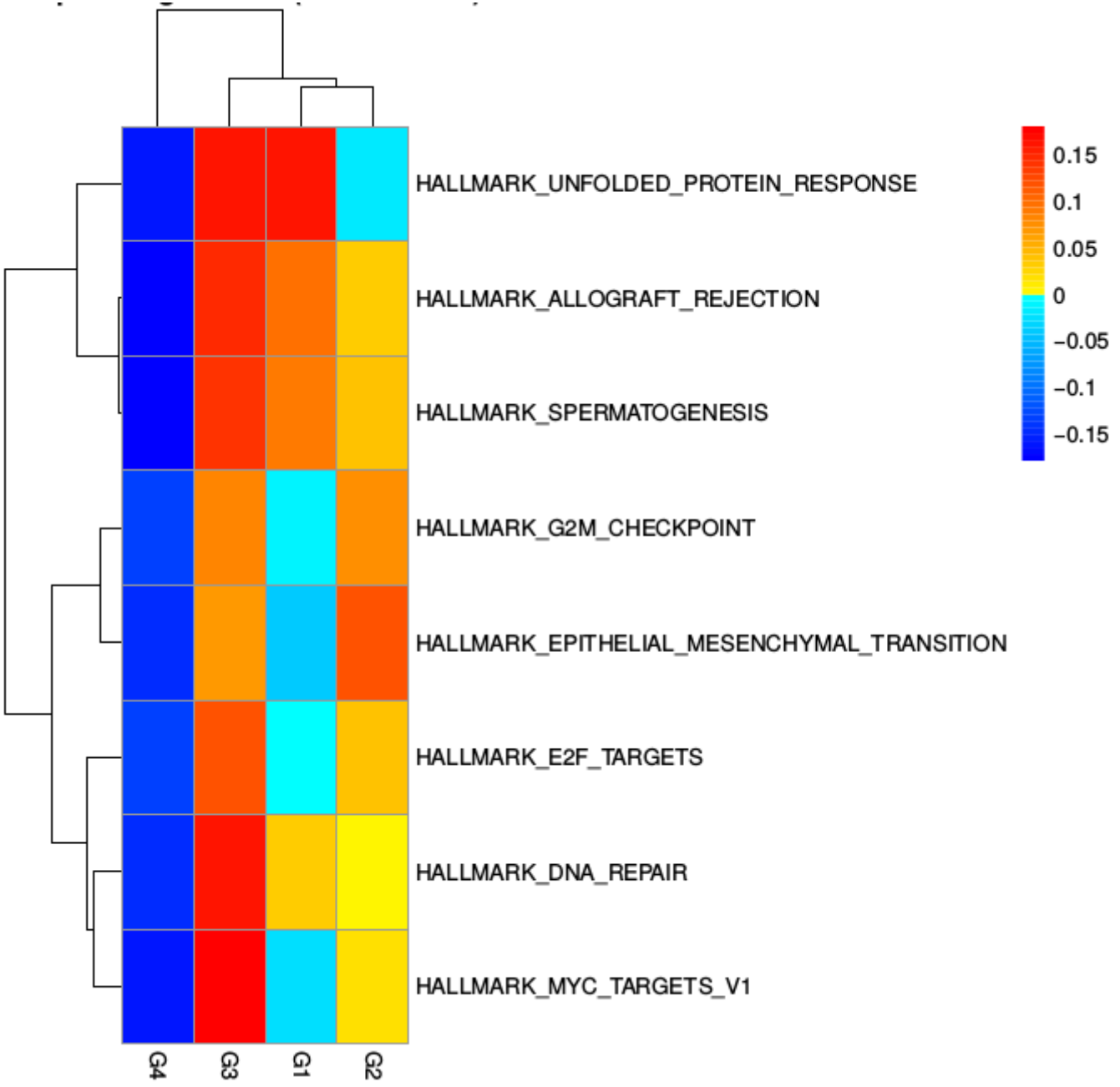
-> Aggregating data by Group

Heatmap of module activity per sample



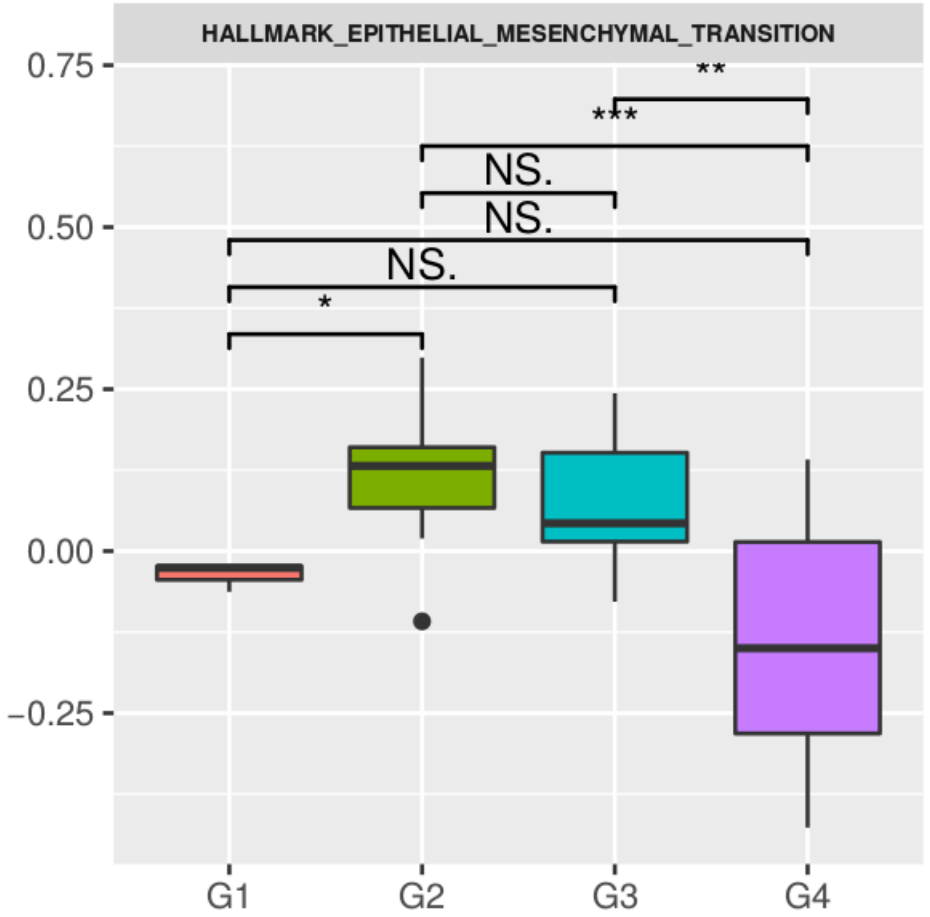
-> EMT active in G2 samples
-> MYC targets activated in G3 samples

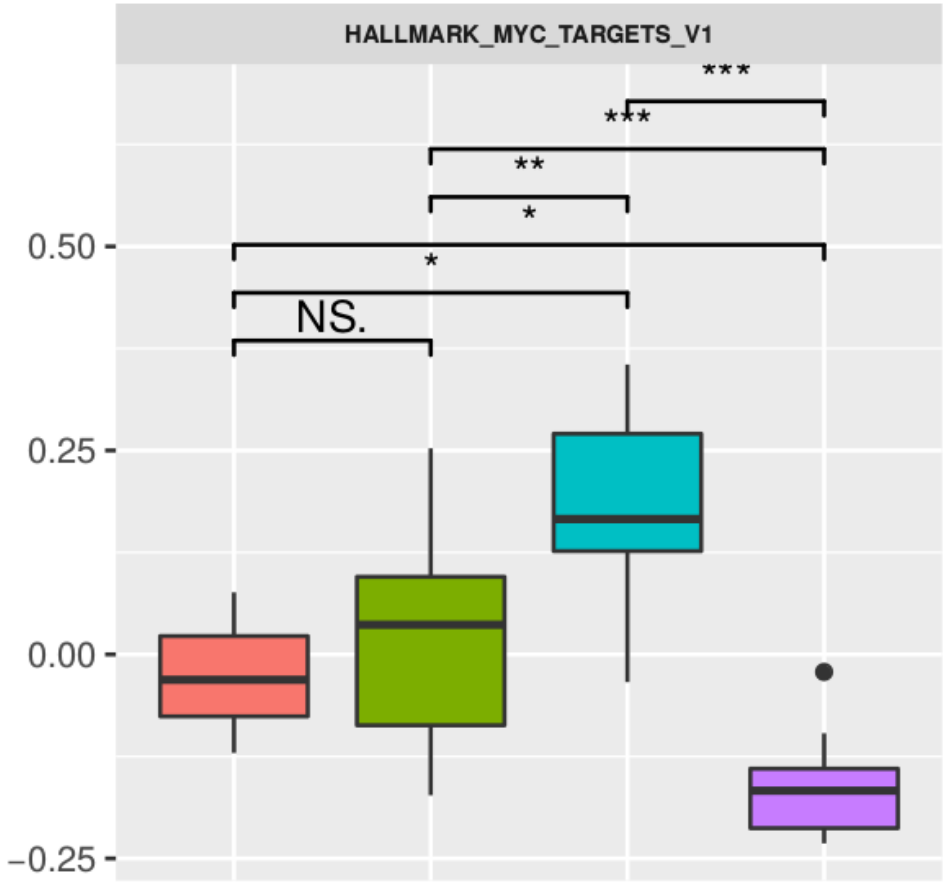
Heatmap of module activity per group



Differential analysis between groups based on module activity

```
CompareAcrossSamples(RomaData = Data.hall,
                     Selected = SelectGeneSets(RomaData = Data.hall,
                                               VarThr = 1e-05,
                                               VarMode = "Wil",
                                               VarType = "Over"),
                     Groups = Group)
```





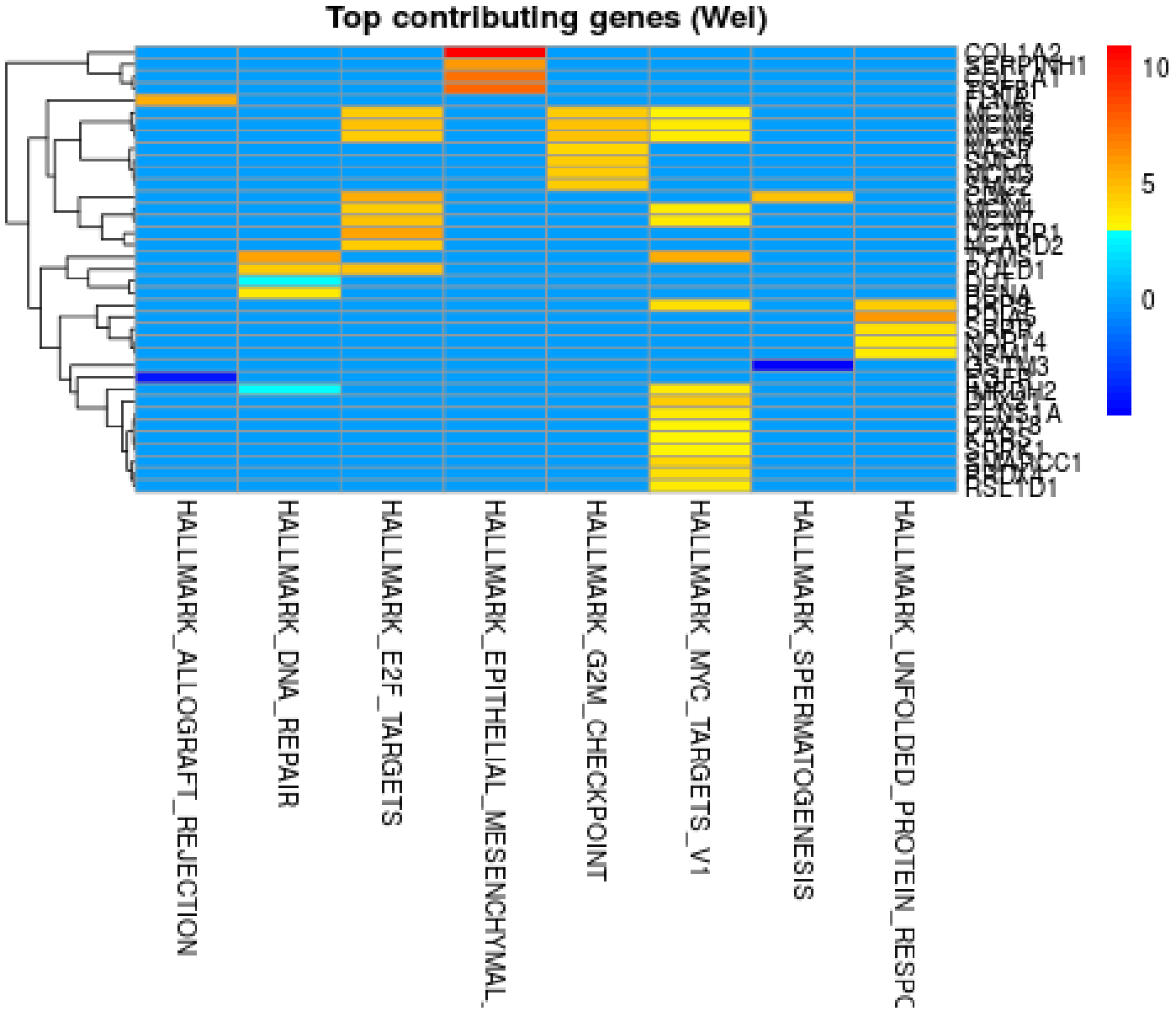
institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

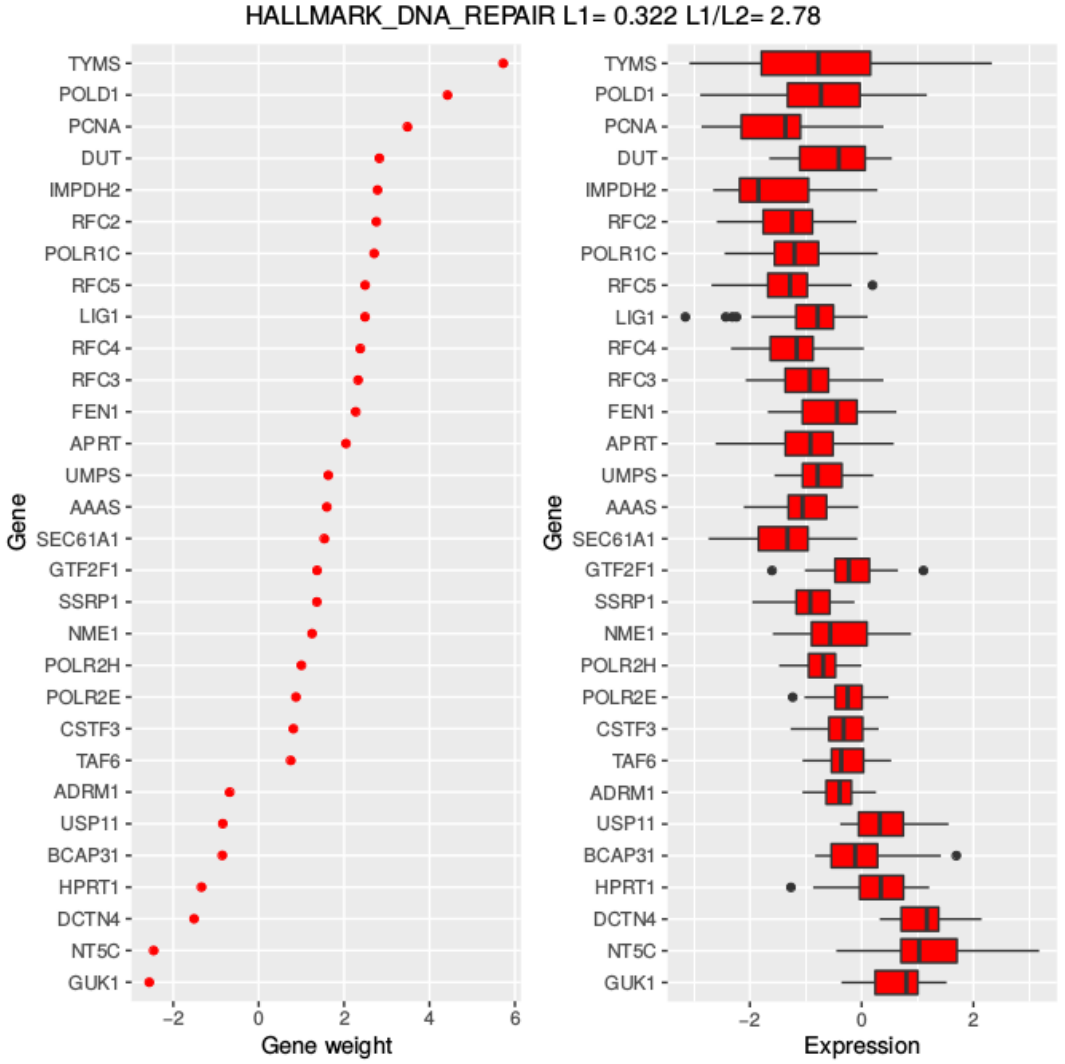
Heatmap of the most contributing genes for significant modules




```
GeneMat <- GetTopContrib(Data.hall,
                Selected = SelectGeneSets(RomaData = Data.hall,
                                VarThr = 1e-5, VarMode = "Wil", VarType = "Over"),
                nGenes = .1, OrderType = "Abs", Mode = "Wei", Plot = TRUE)
```

Visualize the most contributing genes for a given module

```
PlotGeneWeight(RomaData = Data.hall, PlotGenes = 30,
        ExpressionMatrix = expr, LogExpression = FALSE,
        Selected = SelectGeneSets(RomaData = Data.hall,
                VarThr = 1e-5, VarMode = "Wil", VarType = "Over"),
        PlotWeigthSign = TRUE)
```



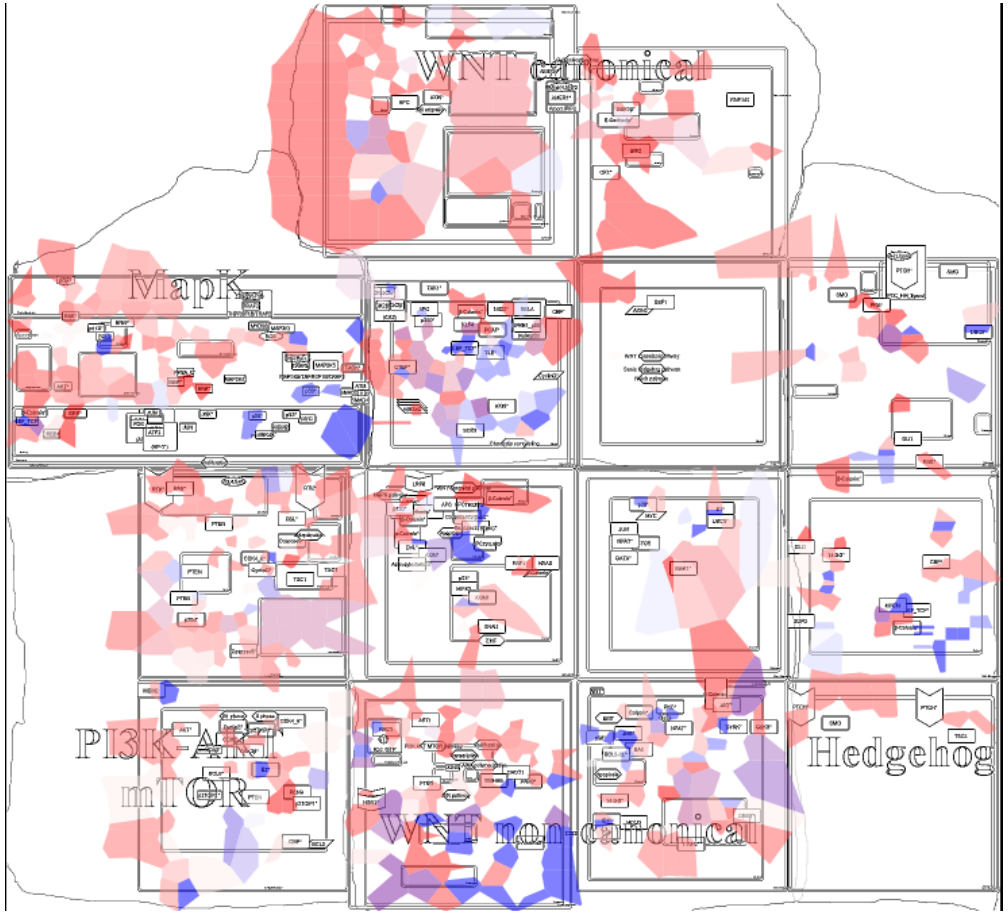
institutCurie

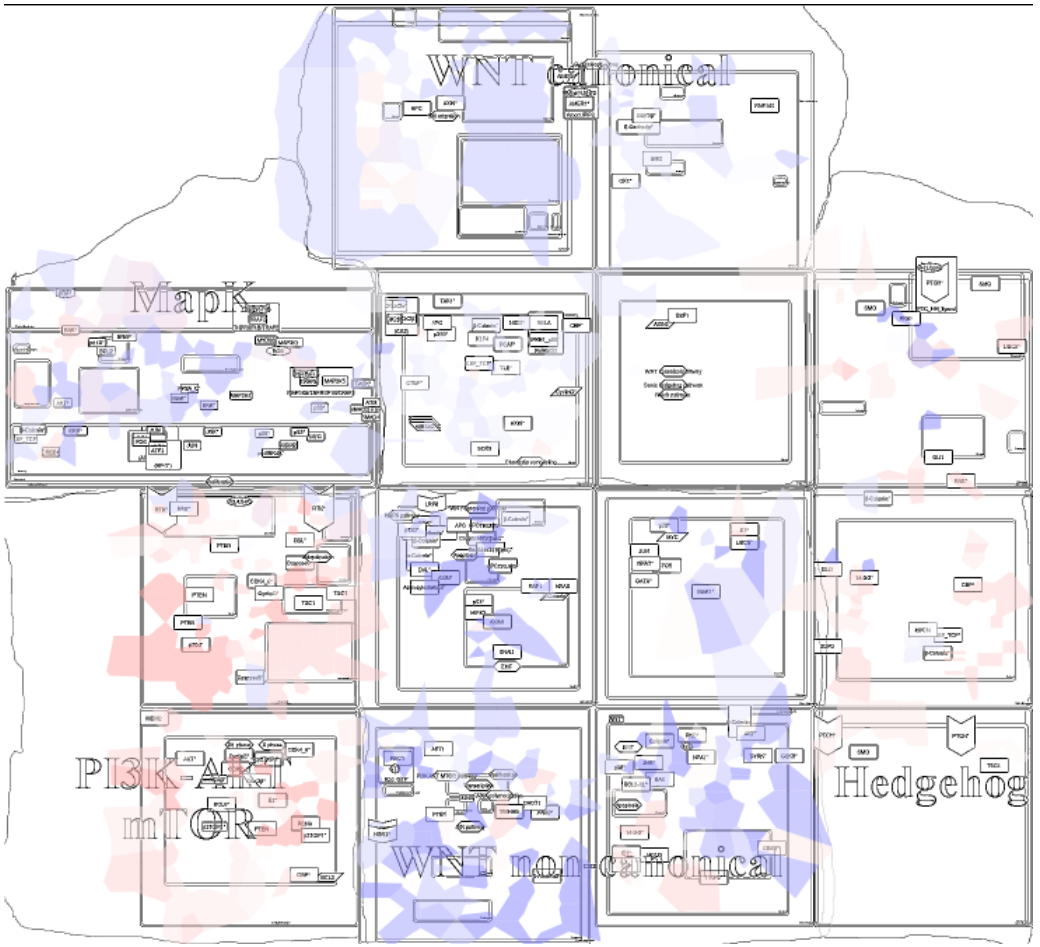
PSL RESEARCH UNIVERSITY

Visualization of ROMA scores on ACSN maps

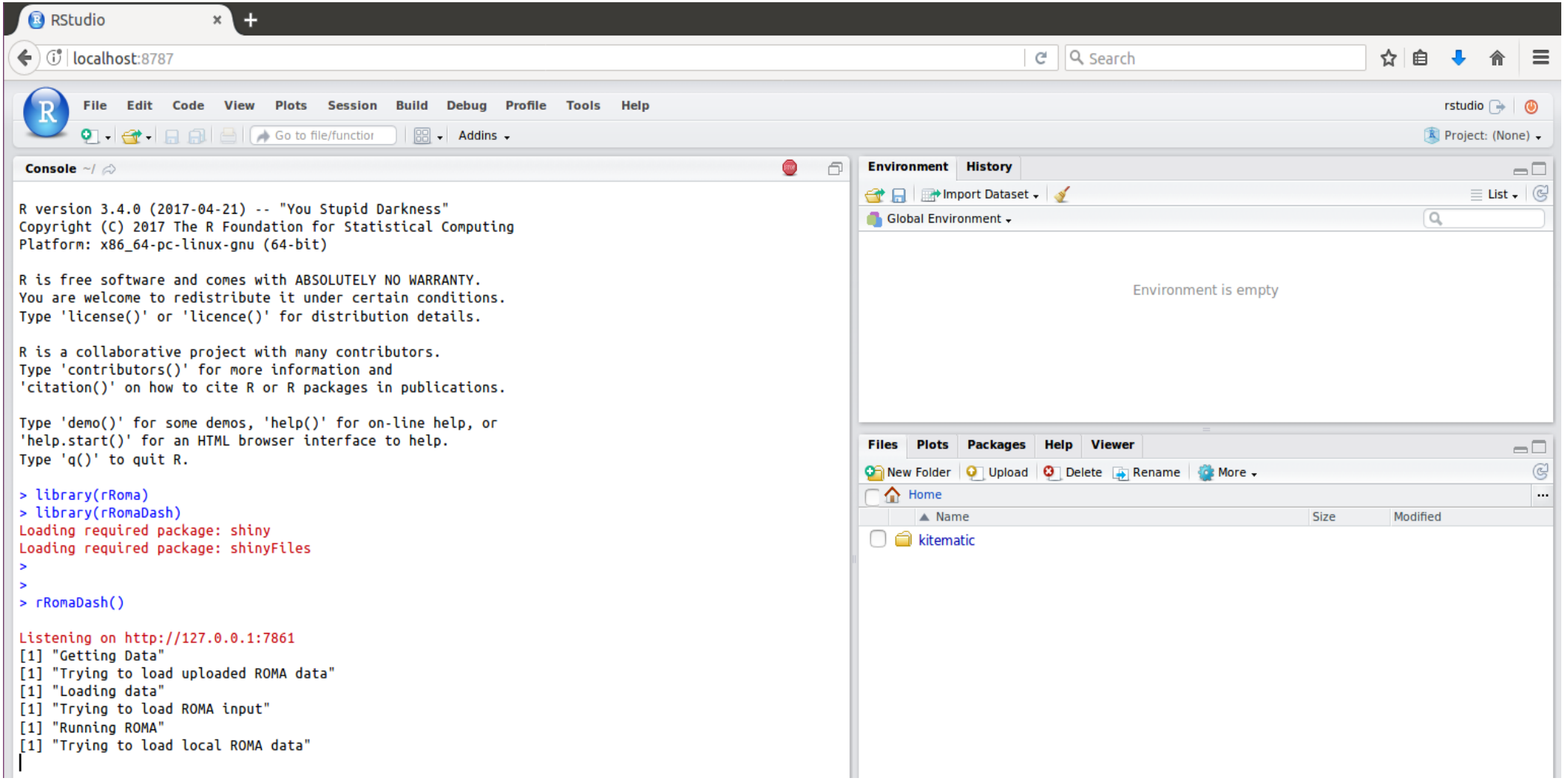
-> Testing gene sets from ACSN maps

-> Visualizing ROMA scores by Group (creating group-specific maps)





Launch rROMA interface : rROMADash()



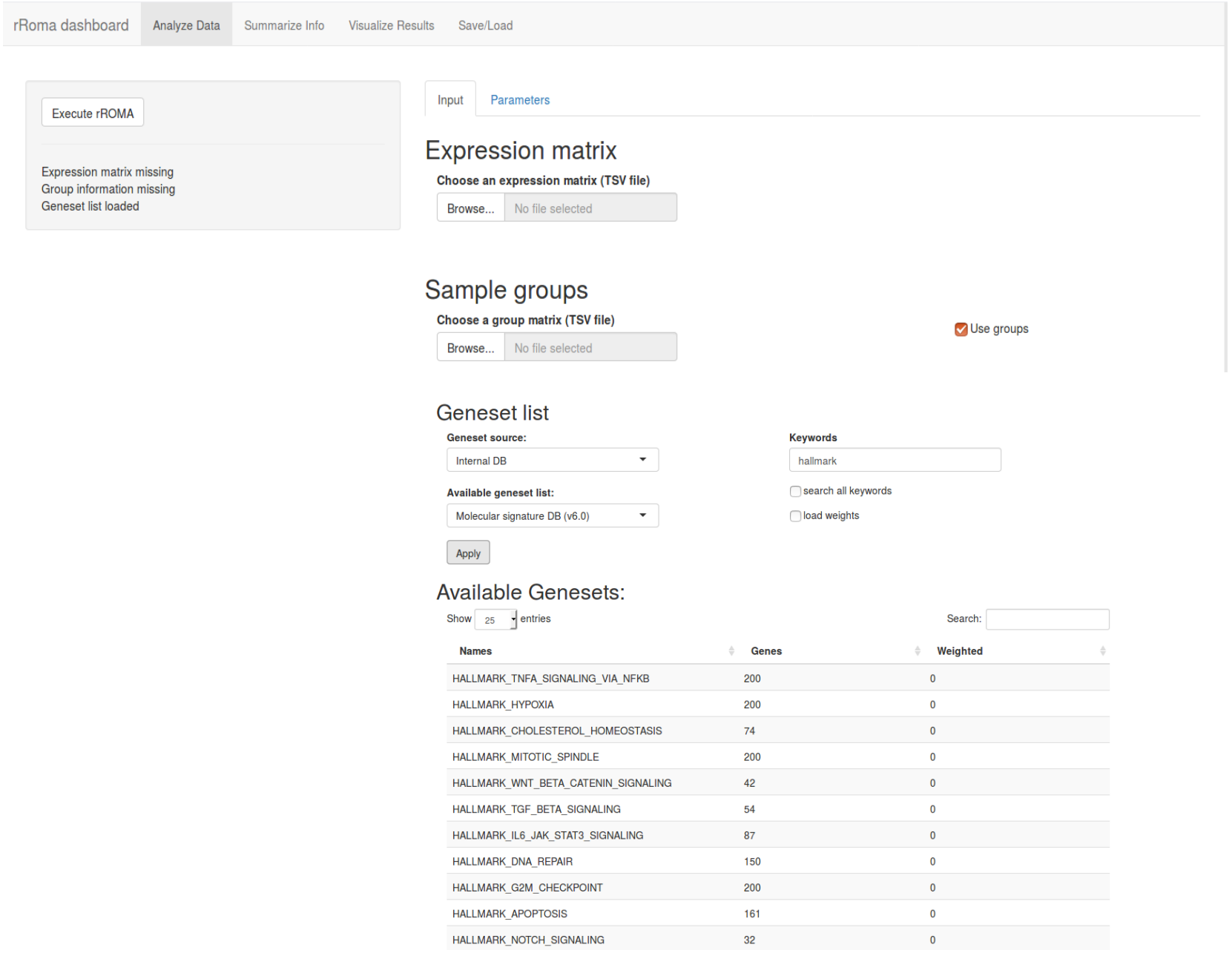
institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Execute rRoma : load data



Execute rRoma : set parameters

Base parameters

FixedCenter
FALSE

UseParallel
TRUE

PCSignMode
CorrelateAllWeightsByGene

nCores
7

nSamples
100

ClusType
PSOCK

GeneOutThr
5

Advanced parameters

UseWeigths
FALSE

SampleFilter
TRUE

MinGenes
10

FullSampleInfo
FALSE

ExpFilter
FALSE

MaxGenes
500

centerData
TRUE

MoreInfo
FALSE

ApproxSamples
5

GeneSelMode
All

GeneOutDetection
L1OutExpOut

PCSignThr
NULL

Ncomp
5

OutGeneNumber
5

CorMethod
pearson

DefaultWeight
1

OutGeneSpace
NULL

institutCurie

PSL RESEARCH UNIVERSITY

Acknowledgments

Computational Systems Biology of Cancer group

Emmanuel Barillot

Luca Albergante
Jonas Béal
Laurence Calzone
Laura Cantini
Urszula Czerwinska
Mihaly Koltai
Maria Kondratova

Loredana Martignetti
Inna Kuperstein
Gaelle Letort
Christine Lonjou
Cristóbal Monraz
Andrei Zinovyev

Resources

https://github.com/sysbio-curie/rRoma

https://github.com/sysbio-curie/rRomaDash

https://github.com/sysbio-curie/Roma

https://github.com/sysbio-curie/Roma_tutorial