

Multi-Omics data integration with the mixOmics package

Sébastien Déjean

www.math.univ-toulouse.fr/~sdejean

GT Biopuces

10 février 2021



Outline

- 1 **Introduction:** interdisciplinarity, data integration, answer a question
- 2 **Tool:** mixOmics R package, workflow
- 3 **Methods:** PCA, extension to integration problems, sparsity, multilevel, vertical integration
- 4 **Examples:** liver toxicity, Wallomics

1 Interdisciplinarity

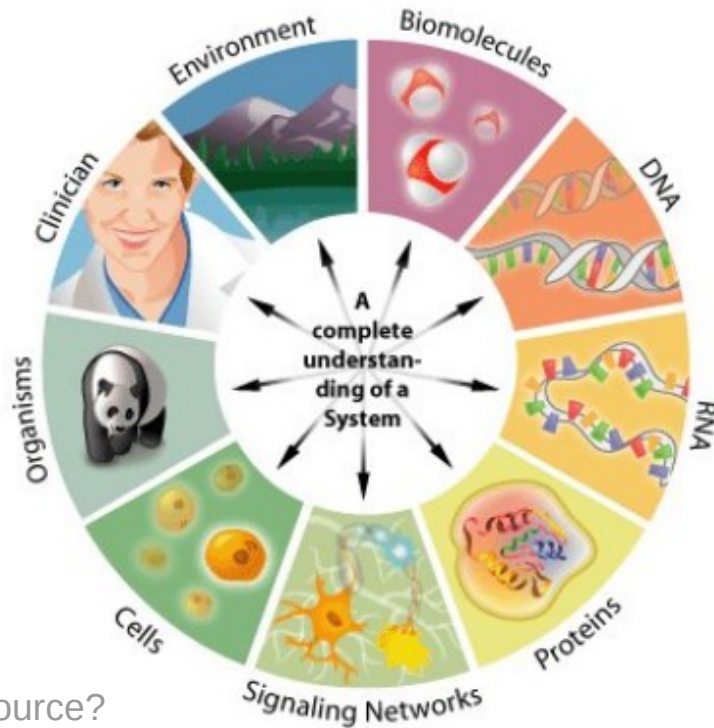
*The biological sciences are **today** in the process of changing from being primarily descriptive **to being very much quantitative**. As a result, biologists find themselves **confronted more and more with large amounts of numerical data** [...]. But the mere collecting and recording of data achieve nothing; having been collected, they must be **investigated to see what information may be contained concerning the biological problem at hand**. [...]*

*Frequently, however, biologists have to subject their data to more complex calculations, requiring procedures that **involve mathematical details beyond their general experience**. In order to carry out the mathematics the biologist in this situation must either **learn the procedures himself**, or at least **learn something of the language of mathematics**, that he may **communicate satisfactorily with the mathematician** whose aid he enlists.*

S.R Searle (**1966**)

Matrix Algebra for the biological sciences

1 Data integration



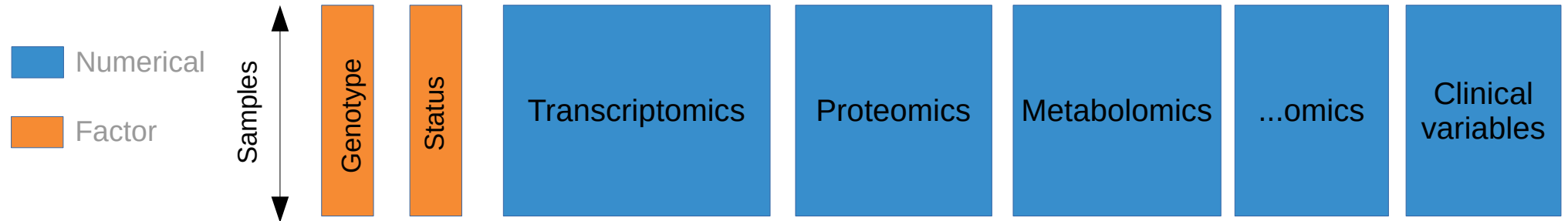
Source?

Generally, data integration can be defined as the process of combining data residing in diverse sources to provide users with a comprehensive view of such data. There is no universal approach to data integration, and many techniques are still evolving.

From Schneider, M. V., & Jimenez, R. C. (2012). Teaching the Fundamentals of Biological Data Integration Using Classroom Games. PLoS Computational Biology, 8(12)

1 Statistical data integration

Analyse simultaneously several datasets to extract knowledge unreachable when considering each dataset separately



1 Answer a question

THE FUTURE OF DATA ANALYSIS¹

BY JOHN W. TUKEY

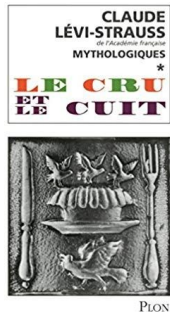
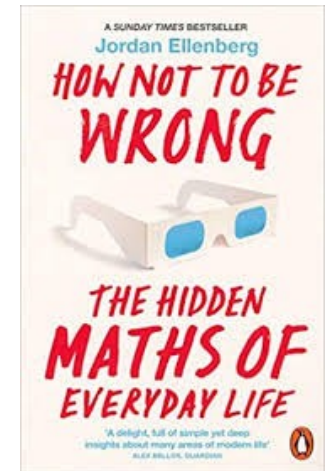
Princeton University and Bell Telephone Laboratories

Received July 1, 1961.

¹ Prepared in part in connection with research sponsored by the Army Research Office through Contract DA36-034-ORD-2297 with Princeton University. Reproduction in whole or part is permitted for any purpose of the United States Government.

*Far better an approximate answer to **the right question** [...], than an exact answer to the wrong question [...].*

*[...] in order to give a sensible answer, you need to know more than just numbers [...]. It's **only after you've started to formulate these questions** that you take out the calculator. But **at that point the real mental work is already finished**. Dividing one number by another is mere computation; figuring out what you should divide by what is mathematics.*



*Le savant n'est pas l'homme qui fournit les vraies réponses; c'est celui qui pose les **vraies questions**.*

C. Lévi-Strauss. Le Cru et le Cuit (1964)

2 The 'Calculator'



- Package for the **R software** r-project.org
- **Born** in Toulouse, France, in 2009
- **Team leader:** Kim-Anh Lê Cao, Melbourne Integrative Genomics, University of Melbourne lecao-lab.science.unimelb.edu.au
- **Freely available** on the repository Bioconductor: bioconductor.org/packages/release/bioc/html/mixOmics.html
- **Web site with tutorial and case studies:** www.mixomics.org
- **Forum:** mixomics-users.discourse.group

2 The mixOmics facebook

- Core team



Sébastien Déjean,
Kim-Anh Lê Cao,
Ignacio Gonzalez,
Florian Rohart

- Key developers / contributors



Benoit Gautier

Al J Abadi



Xin-Yi Chua

François Bartolo



Amrit Singh

Casey Shanon



- Tutors / teachers contributors



Olivier Chapleur



Eva Yiwen Wang



Laëtitia Cardona



David Rengel



Yannick Lippi



Jérôme Mariette

- Many users and trainees



2 mixOmics: key figures

- **>555K** total download since 2009 (CRAN + Bioconductor)
- **>600** attendees for workshops organised since 2014
- **>1 000** citations of the article *mixOmics: an R package for `omics selection and multiple data integration* (Google Scholar, November, 22th)

2 mixOmics workflow

1) Run a method: `pca()`, `spca()`, `pls()`, `spls()`, `plsda()`,
`spplsda()`, `block.pls()`, `block.spls()`, `block.plsda()`, `block.spplsda()`

2) Represent individuals: `plotIndiv()`

3) Represent variables: `plotVar()`, `plotLoadings()`,
`cim()`, `network()`

3 Methods

- Principal Component Analysis
- Multi-blocks methods
- Sparsity
- Multilevel
- Vertical integration (multi-groups methods)

3 Overview of statistical methods available in mixOmics

- Multivariate unsupervised
Principal Components Analysis (PCA)



- Multivariate supervised
*Projection to Latent Structure
Discriminant Analysis (PLS-DA)*



- Multi-block unsupervised
*Canonical Correlation Analysis (CCA) or
PLS (2 blocks), Generalized CCA (>2 blocks)*



- Multi-block supervised
*Generalized Canonical Correlation
Discriminant Analysis (GCC-DA)*



3 Understand PCA

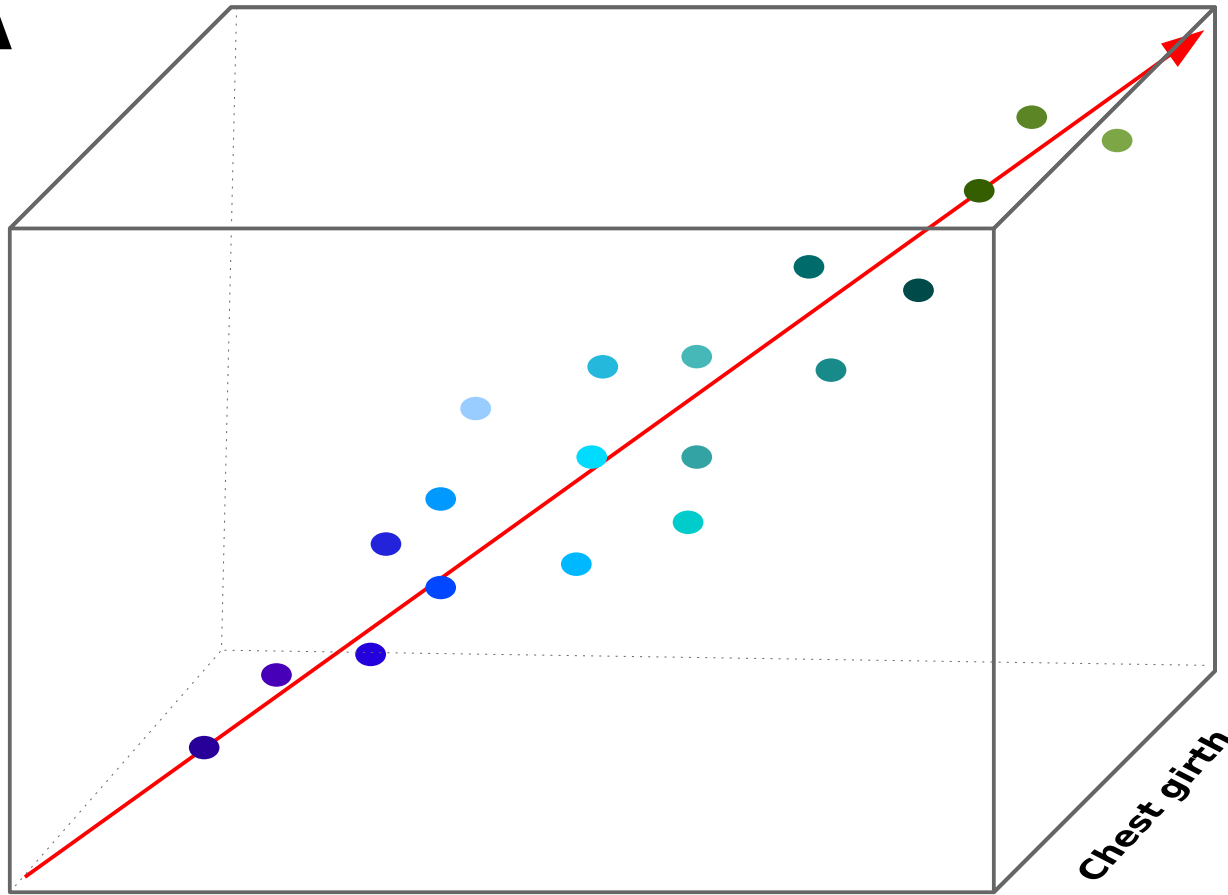
Teasing: would you use a cubic box to pack a fishing rod?



3 PCA

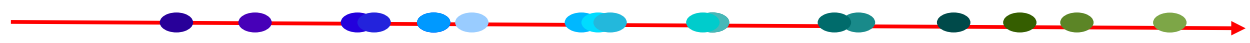


Shoulder girth



Waist girth

Chest girth



1st Principal Component:
«beefyness»

3 A toy example

- 20 individuals
- 5 variables

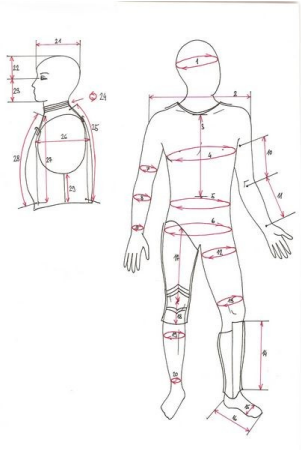
s.g : shoulder girth (cm)

c.g : chest girth (cm)

w.g : waist girth (cm)

w : weight (kg)

h : height (cm)



Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

3 First computations

Bivariate analysis

Raw data

Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

Covariance matrix

	s.g	c.g	w.g	w	h
s.g	68.6	37.7	28.1	55.3	61.2
c.g	37.7	37.5	33.9	45.7	32.4
w.g	28.1	33.9	50.8	56.6	27.7
w	55.3	45.7	56.6	85.7	59.5
h	61.2	32.4	27.7	59.5	109.3

Pearson correlation matrix

	s.g	c.g	w.g	w	h
s.g	1.0	0.7	0.5	0.7	0.7
c.g	0.7	1.0	0.8	0.8	0.5
w.g	0.5	0.8	1.0	0.9	0.4
w	0.7	0.8	0.9	1.0	0.6
h	0.7	0.5	0.4	0.6	1.0

Univariate analysis

Mean	108.1	94.2	75.3	70.6	174.4
Variance	68.6	37.5	50.8	85.7	109.3

351.9 represents the quantity of information contained in the data.

$$68.6 + 37.5 + 50.8 + 85.7 + 109.3 = 351.9$$

3 The core of PCA

Coefficients of linear combination (or loadings)

	PC1	PC2	PC3	PC4	PC5
shoulder.g	0.45	-0.16	0.78	-0.18	0.36
chest.g	0.32	0.25	0.26	0.72	-0.49
waist.g	0.34	0.53	-0.33	0.24	0.66
weight	0.54	0.36	-0.17	-0.60	-0.44
height	0.54	-0.70	-0.43	0.17	0.02

PC1 = 0.45*shoulder.g + 0.32*chest.g + 0.34*waist.g + 0.54*weight + 0.54*height

PC2 = -0.16*shoulder.g + 0.25*chest.g + 0.53*waist.g + 0.36*weight - 0.70*height

...

Q: Where do these coefficients come from?

A: Matrix algebra, eigen decomposition of the covariance matrix or singular value decomposition of the initial matrix

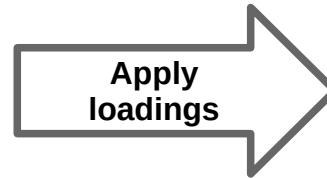
3 Around the core

Centered data

$$\text{Ex: } -6.50 = 0.45*(-1.9) + 0.32*(-4.7) + 0.34*(-3.8) + 0.54*(-5) + 0.54*(-0.4)$$

Id	s.g	c.g	w.g	w	h
I1	-1.9	-4.7	-3.8	-5.0	-0.4
I2	2.4	2.8	3.7	1.2	0.9
I3	7.0	3.3	7.9	10.1	19.1
I4	-3.6	2.8	2.5	2.0	12.1
I5	-0.6	3.3	4.7	8.2	12.8
I6	11.7	5.7	7.2	4.2	7.1
I7	15.4	12.7	6.7	15.8	9.6
I8	12.3	8.3	1.5	7.8	10.1
I9	2.9	-3.2	-6.8	-8.6	0.6
I10	11.4	-0.7	2.2	11.0	9.6
I11	-3.1	-5.2	-4.1	-3.3	-4.9
I12	-7.9	-0.1	4.2	4.9	-14.4
I13	-9.0	-3.4	2.6	-2.4	-1.7
I14	-0.5	2.8	-5.8	-9.2	-11.8
I15	-4.1	1.2	10.7	6.2	-16.9
I16	0.3	-2.4	-5.4	1.2	2.1
I17	-8.8	-6.9	-11.8	-15.1	-10.0
I18	-16.2	-16.1	-17.4	-22.0	-13.7
I19	-1.0	-3.3	-3.1	-4.2	-0.4
I20	-7.6	2.9	5.1	-3.3	-10.6

	PC1	PC2	PC3	PC4	PC5
s.g	0.45	-0.16	0.78	-0.18	0.36
c.g	0.32	0.25	0.26	0.72	-0.49
w.g	0.34	0.53	-0.33	0.24	0.66
w	0.54	0.36	-0.17	-0.60	-0.44
h	0.54	-0.70	-0.43	0.17	0.02



	PC1	PC2	PC3	PC4	PC5
I1	-6.50	-4.48	-0.37	-1.03	1.27
I2	4.40	2.04	0.81	1.87	1.38
I3	22.66	-5.94	-6.18	0.11	1.97
I4	7.78	-5.24	-8.38	4.10	-1.74
I5	13.73	-2.67	-8.02	0.82	-2.15
I6	15.67	-0.15	4.49	2.33	4.40
I7	26.99	3.19	6.29	0.04	-3.08
I8	18.41	-3.43	5.63	1.09	-1.96
I9	-6.25	-8.48	4.97	0.79	1.86
I10	16.78	-3.67	1.99	-7.08	1.22
I11	-8.83	-0.78	0.28	-3.02	0.07
I12	-7.28	15.41	-2.31	-3.00	-2.35
I13	-6.45	2.25	-7.60	0.95	1.15
I14	-12.51	2.68	8.91	4.27	-1.53
I15	-3.65	20.76	-0.30	-2.45	1.99
I16	-0.63	-4.62	0.34	-3.46	-2.80
I17	-23.61	-5.07	2.20	1.19	-1.15
I18	-37.50	-9.07	-1.33	-1.89	-0.02
I19	-4.98	-3.61	0.33	-0.50	1.02
I20	-8.24	10.89	-1.74	4.86	0.44

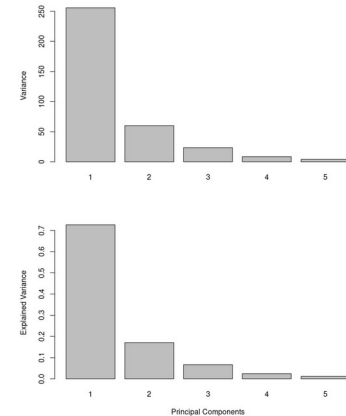
255.7 is the highest variance we can obtain with a linear combination of the initial variables.

Mean 0 0 0 0 0
 Var. **255.7** 60.2 23.5 8.6 4.0 = **351.9**

3 Graphical outputs (1/3)

	PC1	PC2	PC3	PC4	PC5
Variance	255.7	60.2	23.5	8.6	4.0
% variance	72.6	17.1	6.7	2.4	1.1

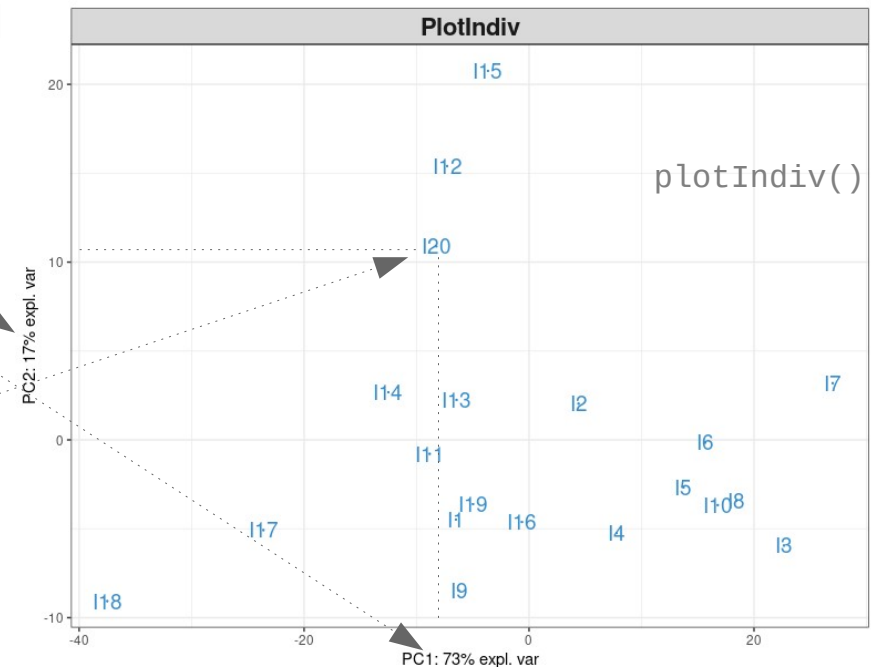
Screeplot



plot()

	PC1	PC2	PC3	PC4	PC5
I1	-6.50	-4.48	-0.37	-1.03	1.27
I2	4.40	2.04	0.81	1.87	1.38
I3	22.66	-5.94	-6.18	0.11	1.97
I4	7.78	-5.24	-8.38	4.10	-1.74
I5	13.73	-2.67	-8.02	0.82	-2.15
I6	15.67	-0.15	4.49	2.33	4.40
I7	26.99	3.19	6.29	0.04	-3.08
I8	18.41	-3.43	5.63	1.09	-1.96
I9	-6.25	-8.48	4.97	0.79	1.86
I10	16.78	-3.67	1.99	-7.08	1.22
I11	-8.83	-0.78	0.28	-3.02	0.07
I12	-7.28	15.41	-2.31	-3.00	-2.35
I13	-6.45	2.25	-7.60	0.95	1.15
I14	-12.51	2.68	8.91	4.27	-1.53
I15	-3.65	20.76	-0.30	-2.45	1.99
I16	-0.63	-4.62	0.34	-3.46	-2.80
I17	-23.61	-5.07	2.20	1.19	-1.15
I18	-37.50	-9.07	-1.33	-1.89	-0.02
I19	-4.98	-3.61	0.33	-0.50	1.02
I20	-8.24	10.89	-1.74	4.86	0.44

Individual plot



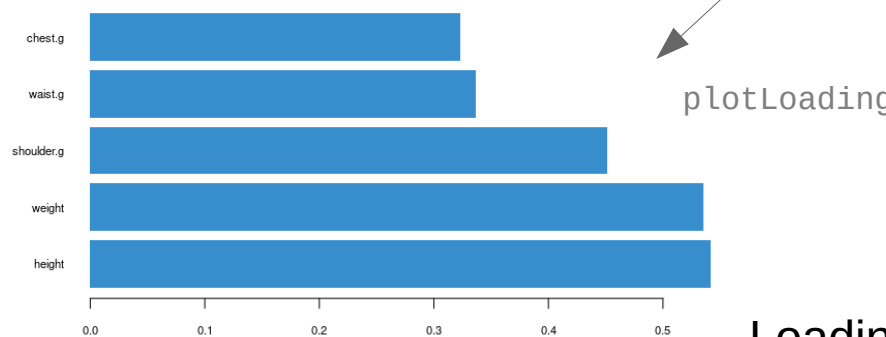
3 Graphical outputs (2/3)

Loadings

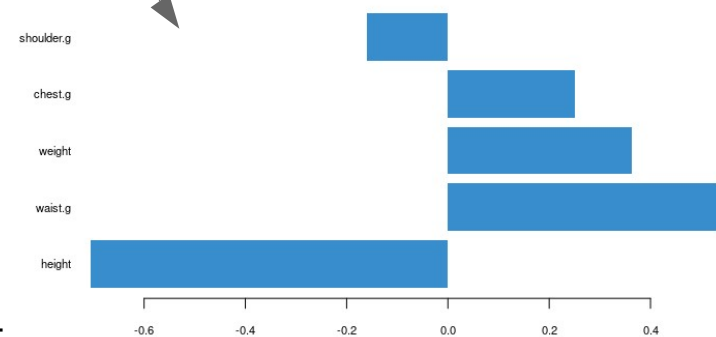
shoulder.g
chest.g
waist.g
weight
height

	PC1	PC2
shoulder.g	0.45	-0.16
chest.g	0.32	0.25
waist.g	0.34	0.53
weight	0.54	0.36
height	0.54	-0.70

Loadings on comp 1



Loadings on comp 2



`plotLoadings()`

Loading plot

3 Graphical outputs (3/3)

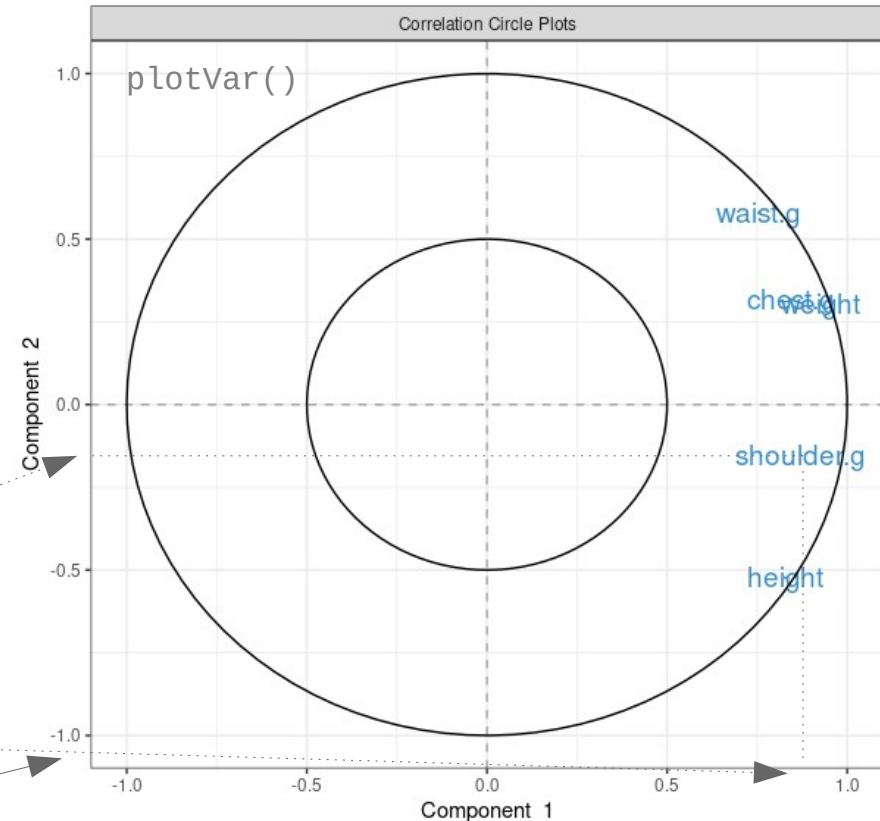
Id	s.g	c.g	w.g	w	h	PC1	PC2	
I1	106.2	89.5	71.5	65.6	174.0	I1	-6.50	-4.48
I2	110.5	97.0	79.0	71.8	175.3	I2	4.40	2.04
I3	115.1	97.5	83.2	80.7	193.5	I3	22.66	-5.94
I4	104.5	97.0	77.8	72.6	186.5	I4	7.78	-5.24
I5	107.5	97.5	80.0	78.8	187.2	I5	13.73	-2.67
I6	119.8	99.9	82.5	74.8	181.5	I6	15.67	-0.15
I7	123.5	106.9	82.0	86.4	184.0	I7	26.99	3.19
I8	120.4	102.5	76.8	78.4	184.5	I8	18.41	-3.43
I9	111.0	91.0	68.5	62.0	175.0	I9	-6.25	-8.48
I10	119.5	93.5	77.5	81.6	184.0	I10	16.78	-3.67
I11	105.0	89.0	71.2	67.3	169.5	I11	-8.83	-0.78
I12	100.2	94.1	79.6	75.5	160.0	I12	-7.28	15.41
I13	99.1	90.8	77.9	68.2	172.7	I13	-6.45	2.25
I14	107.6	97.0	69.6	61.4	162.6	I14	-12.51	2.68
I15	104.0	95.4	86.0	76.8	157.5	I15	-3.65	20.76
I16	108.4	91.8	69.9	71.8	176.5	I16	-0.63	-4.62
I17	99.3	87.3	63.5	55.5	164.4	I17	-23.61	-5.07
I18	91.9	78.1	57.9	48.6	160.7	I18	-37.50	-9.07
I19	107.1	90.9	72.2	66.4	174.0	I19	-4.98	-3.61
I20	100.5	97.1	80.4	67.3	163.8	I20	-8.24	10.89

cor(s.g, PC1) = 0.87
 cor(s.g, PC2) = 0.15

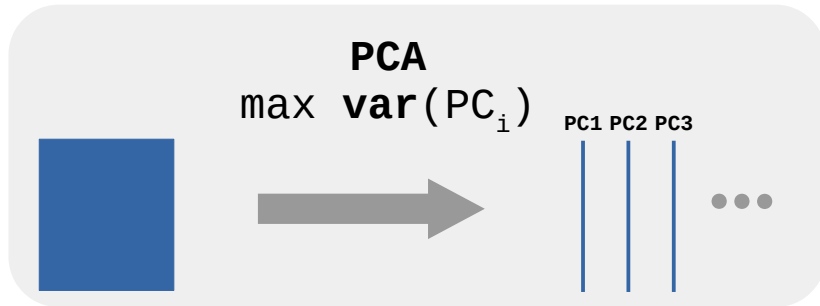
cor(c.g, PC1) = 0.84
 cor(c.g, PC2) = 0.32
 ...

	PC1	PC2
shoulder.g	0.87	-0.15
chest.g	0.84	0.32
waist.g	0.75	0.58
weight	0.92	0.30
height	0.83	-0.52

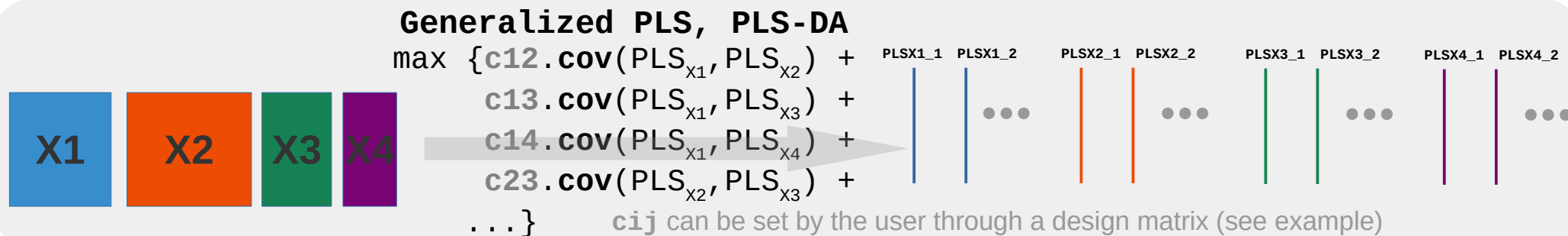
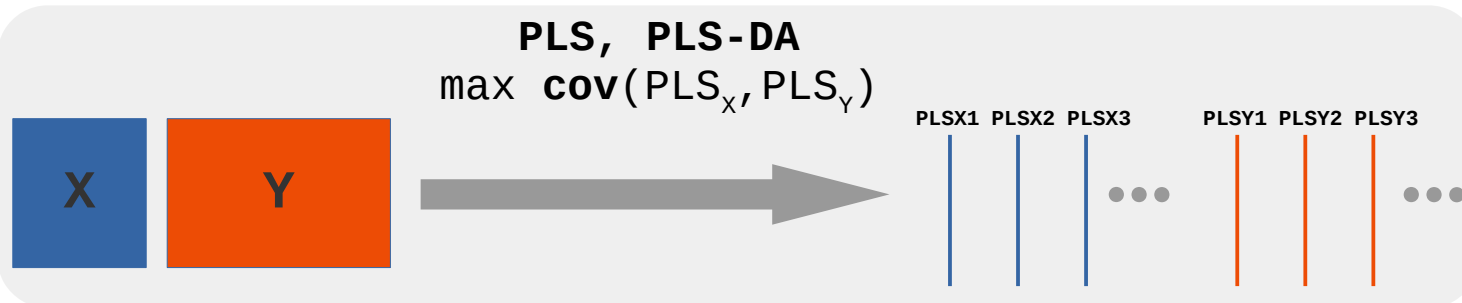
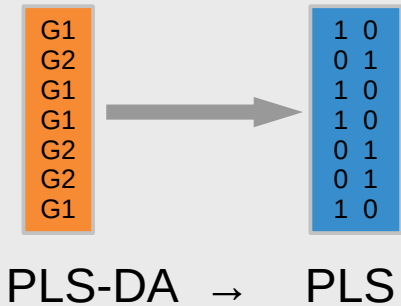
Variable plot



3 Extension to integration problems

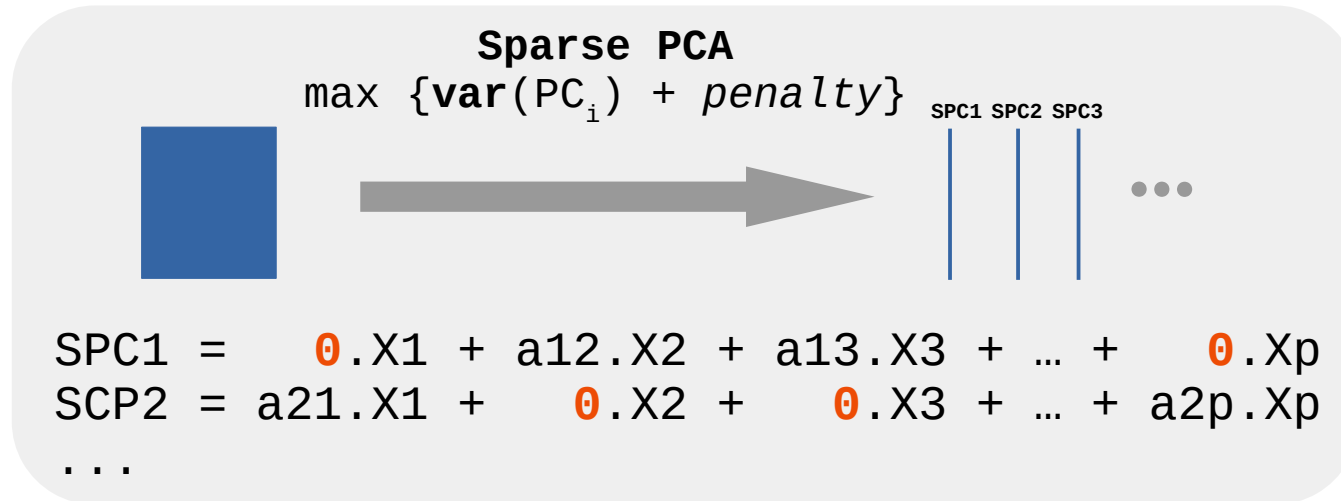


The trick for discriminant analyses: convert a factor into a numeric (dummy) matrix



3 Sparsity

- High throughput experiments: too many variables, noisy or irrelevant depending on the goal aimed
- Some of the variable loadings, among the smallests, are set to 0 thanks to a LASSO (L^1) penalty
- Associated variables are not taken into account when calculating the PCs



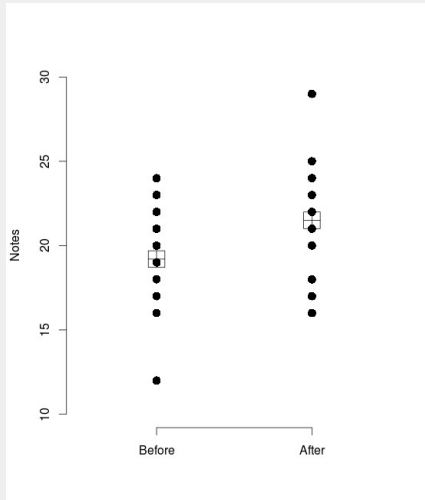
3 Multilevel

- In repeated measures experiments, **the subject variation can be larger than the time/treatment variation**
- Multivariate projection based methods make the assumption that samples are independent of each other
- In univariate analysis we use a **paired** t-test rather than a t-test
- In multivariate analysis we use a **multilevel** approach
- Different sources of variation can be separated (treatment effect within subjects and differences between subjects)

3 Multilevel

Gr.1 Gr.2

18 22
21 25
16 17
22 24
19 18
24 29
17 20
20 23
23 21
12 16



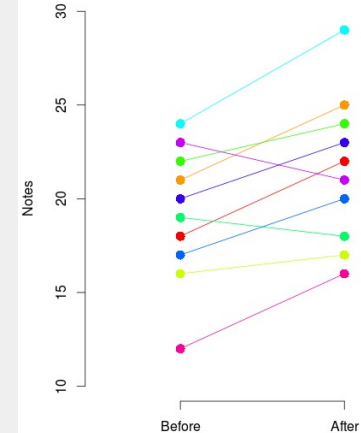
```
> t.test(x,y, paired=FALSE)
Two Sample t-test
```

```
data: x and y
t = -1.3529, df = 18, p-value = 0.1928
alternative hypothesis: true difference
in means
is not equal to 0
95 percent confidence interval:
-5.871567 1.271567
sample estimates:
mean of x mean of y
19.2 21.5
```

Independent data

Before After

Louise	18	22
Léo	21	25
Emma	16	17
Gabriel	22	24
Chloé	19	18
Adam	24	29
Lola	17	20
Timéo	20	23
Inès	23	21
Raphaël	12	16



```
> t.test(x,y, paired=TRUE)
Paired t-test
```

```
data: x and y
t = -3.1461, df = 9, p-value = 0.01181
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-3.953766 -0.646234
sample estimates:
mean of the differences
-2.3
```

Paired data

3 Multilevel

Decomposition of the data into within and between variations

$$X = X_m + X_b + X_w$$

offset term between-sample **within-sample**

- The multilevel approach extracts the **within variation matrix**
- Classical multivariate tools can then be applied on the **within matrix**

3 Multilevel: toy example

3 variables (A, B, C) measured for 10 sujets (1...10) in 2 conditions *control* ou *treatment*.

Raw data set

condition	subject	A	B	C
control	1	20	10	20
control	2	18	12	17
control	3	16	15	14
control	4	14	16	11
control	5	10	2	8
control	6	9	3	5
control	7	7	7	2
control	8	7	7	8
control	9	3	9	14
control	10	2	9	17
treatment	1	21	12	20
treatment	2	21	14	17
treatment	3	17	17	14
treatment	4	17	18	11
treatment	5	11	4	8
treatment	6	12	5	5
treatment	7	8	9	2
treatment	8	10	9	8
treatment	9	4	11	14
treatment	10	5	11	17

Between-subject matrix

subject	A	B	C
1	20.5	11	20
2	19.5	13	17
3	16.5	16	14
4	15.5	17	11
5	10.5	3	8
6	10.5	4	5
7	7.5	8	2
8	8.5	8	8
9	3.5	10	14
10	3.5	10	17
1	20.5	11	20
2	19.5	13	17
3	16.5	16	14
4	15.5	17	11
5	10.5	3	8
6	10.5	4	5
7	7.5	8	2
8	8.5	8	8
9	3.5	10	14
10	3.5	10	17

Within-subject matrix

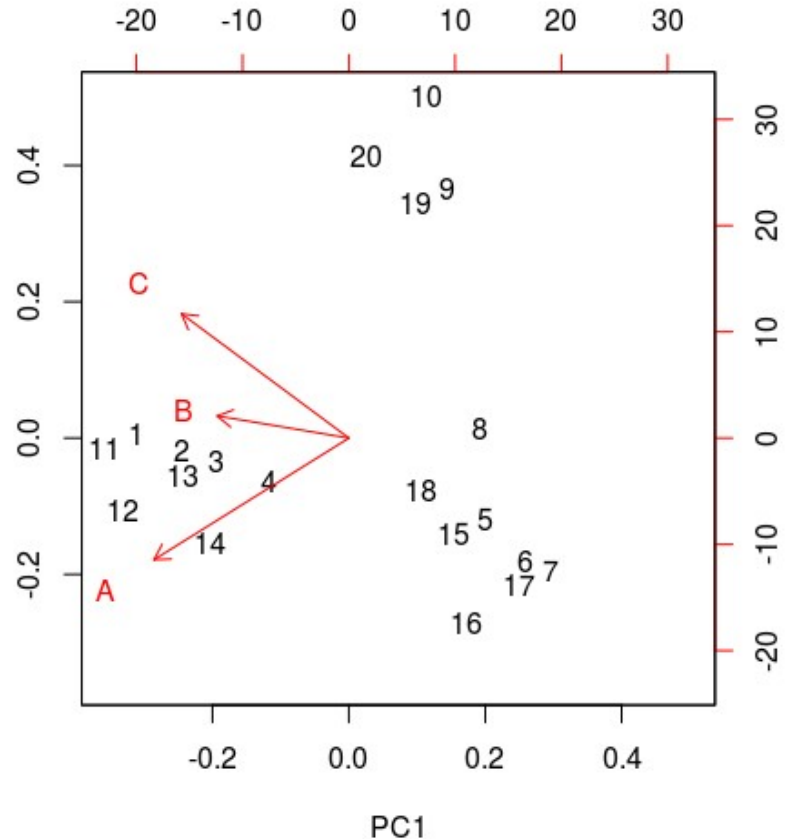
DA	DB	DC
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0

3 Multilevel: toy example

PCA on raw data

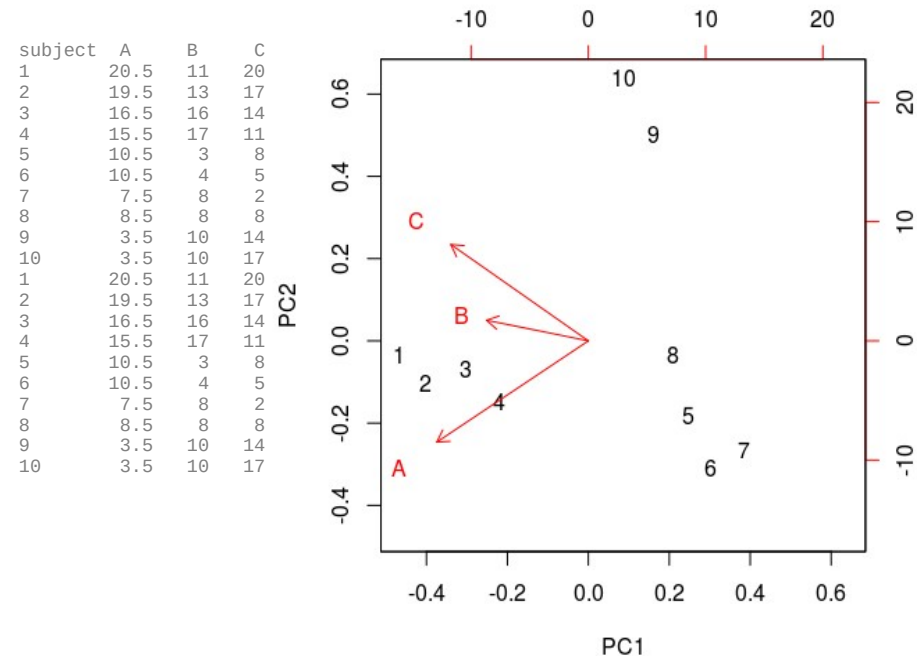
- The main information relies on the close locations of the two measurements made on each subject (1-11, 2-12, ..., 9-19, 10-20)
- No treatment effect can be observed

condition	subject	A	B	C
control	1	20	10	20
control	2	18	12	17
control	3	16	15	14
control	4	14	16	11
control	5	10	2	8
control	6	9	3	5
control	7	7	7	2
control	8	7	7	8
control	9	3	9	14
control	10	2	9	17
treatment	1	21	12	20
treatment	2	21	14	17
treatment	3	17	17	14
treatment	4	17	18	11
treatment	5	11	4	8
treatment	6	12	5	5
treatment	7	8	9	2
treatment	8	10	9	8
treatment	9	4	11	14
treatment	10	5	11	17



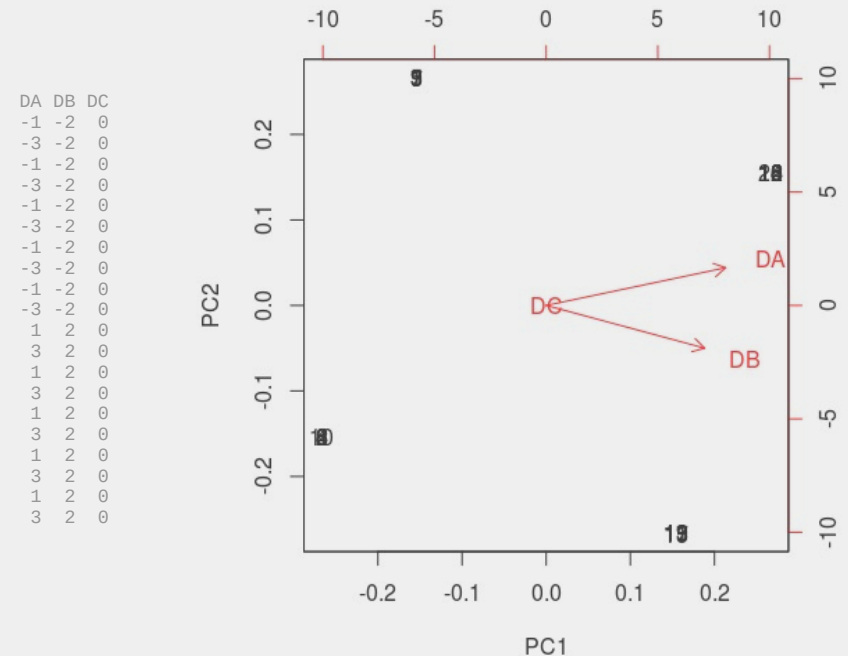
3 Multilevel: toy example

PCA on between matrix



- Nearly the same information as obtained on the raw data
- Because variability between subjects is greater than the variability due to the treatment

PCA on within matrix



- Only 4 distinct points (related to the 4 unique rows in the within matrix)
- Treatment effect clearly appears

3 Multilevel: in practice

```
R> library(mixOmics)
R> pca(MyData
      multilevel = subject)
R> spca(MyData
        multilevel = subject)
R> plsda(MyData, OutCome,
         multilevel = subject)
R> ...
```

- Case study:

mixomics.org/case-studies/multilevel-vac18/

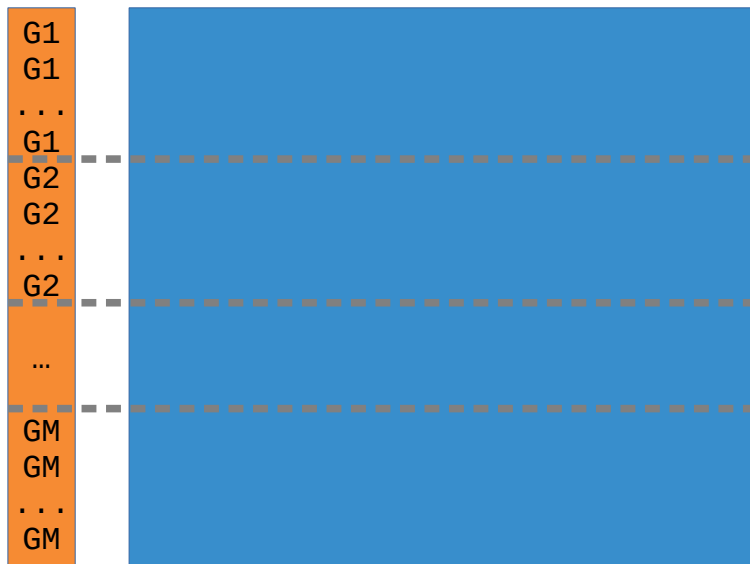
3 Vertical (P-) integration: multi-group PCA

- *Setting: the same variables measured on individuals portioned into several groups*
- *The same setting as in discriminant analysis **but** the main aim herein is to investigate the relationships among individuals within the various groups*



A. Eslami, E.M Qannari, A. Kohler, S. Bougeard (2013). Analyses factorielles de données structurées en groupes d'individus. Journal de la SfdS, vol. 154(3). journal-sfds.fr/article/view/208

www.rocq.inria.fr/axis/modulad///sda11/HCSDA11-Qannari.PDF



***Ask the
right
question!***

3 Vertical integration: mgPCA

How to investigate the relationships among individuals within the various groups?

- **Perform PCA on each group separately**

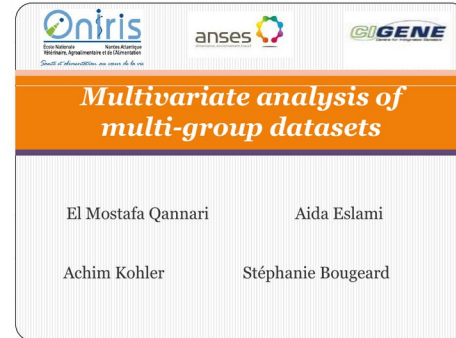
→ Too many parameters (stability and interpretation problems)

- **Perform PCA on the concatenated dataset**




→ The total variance recovered by the principal components mix up both the between and within groups variances

- **Multi-group PCA**

→ Perform PCA on the concatenated dataset **after centering by group**

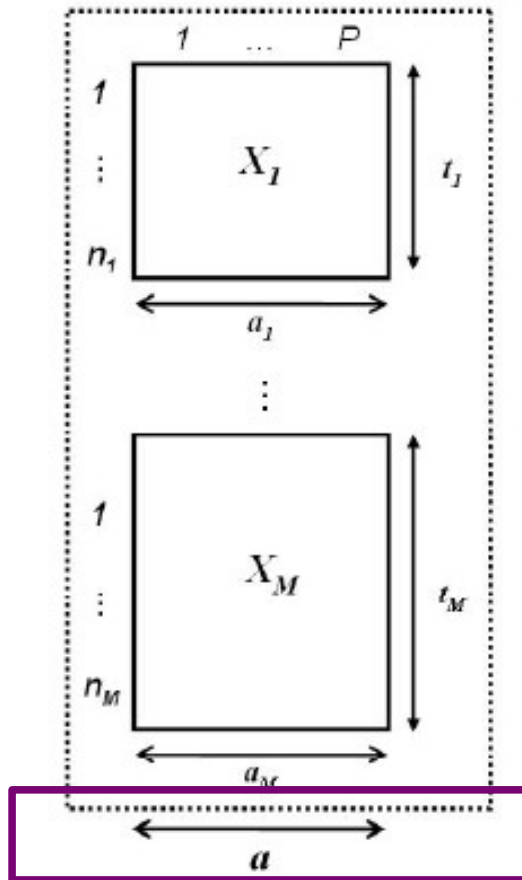


3 Vertical integration: mgPCA

Multivariate analysis of multi-group datasets

El Mostafa Qannari Aida Eslami
 Achim Kohler Stéphanie Bougeard



A vector of loadings associated with X_m is given by:

$$a_m = X_m^T t_m$$

Relationship between a (common vector of loadings) and λ_m (specific variance to group m):

$$\lambda_m = \text{var}(X_m a) = a^T V_m a$$

- Maximize:

$$\sum_m n_m \text{var}(t_m) \quad \text{with } t_m = X_m a \quad \text{and} \quad \|a\| = 1$$

- Find a common vector of loadings, a , so as to maximize:

$$\sum_m \langle a_m, a \rangle^2 \quad \text{with } a_m = X_m^T t_m$$

$$\|a_m\| = \|a\| = 1$$

a : vector of common loadings \rightarrow the same variable plot for every group



3 Vertical integration

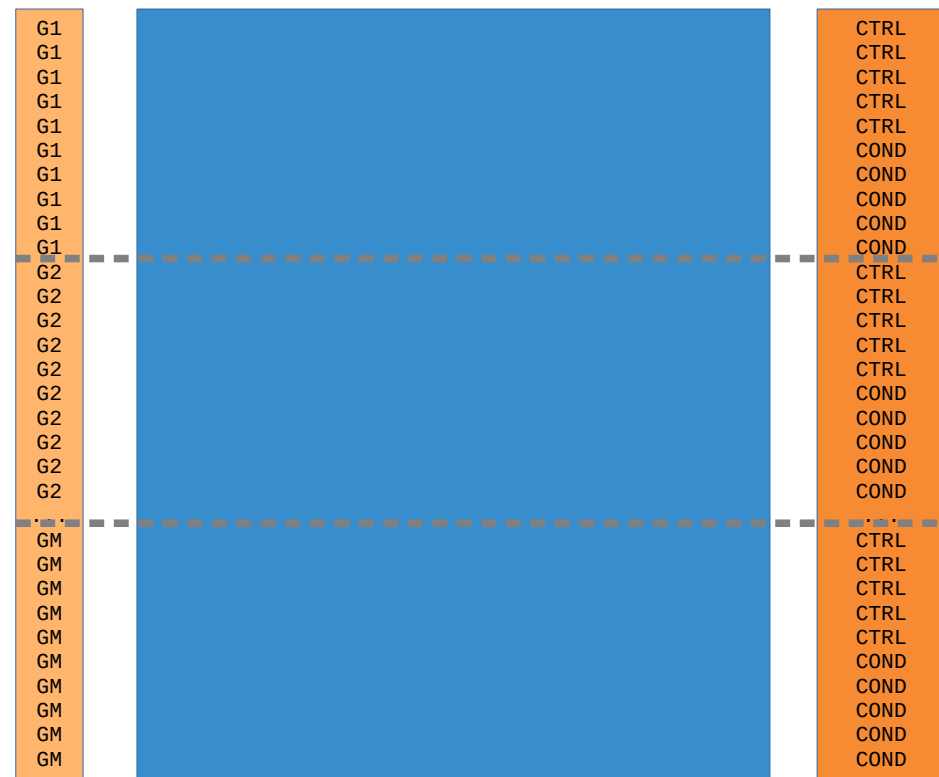
MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms *BMC Bioinformatics* 18:128.

Florian Rohart¹, Aida Eslami², Nicholas Matigian¹, Stéphanie Bougeard³ and Kim-Anh Lê Cao^{1*}

MINT PLS-DA

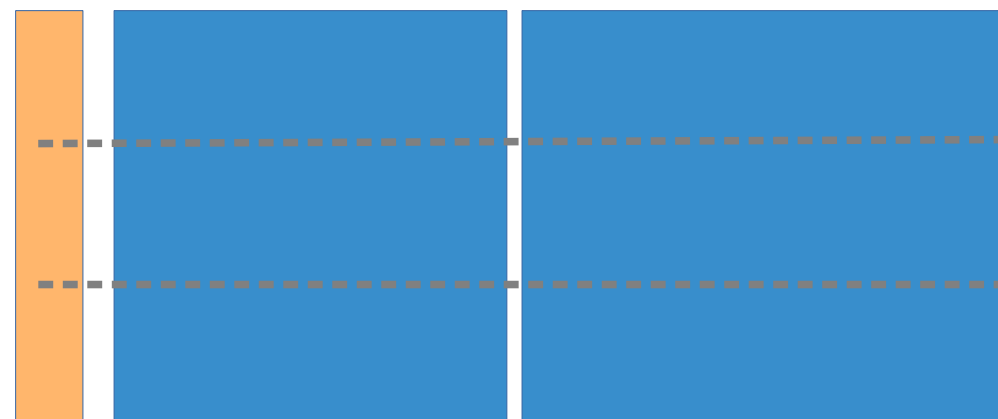
Group

Condition



While PLS-DA ignores the data group structure inherent to each independent study, it can give satisfactory results when the between groups variance is smaller than the within group variance.

MINT PLS



3 Vertical integration

the component. For each dimension $h = 1, \dots, H$ PLS-DA seeks to maximize

$$\max_{\|a_h\|_2=\|b_h\|_2=1} \text{cov}(X_h a_h, Y_h b_h), \quad (1)$$

For each dimension $h = 1, \dots, H$ the MINT algorithm seeks to maximize (m) group index

$$\max_{\|a_h\|_2=\|b_h\|_2=1} \sum_{m=1}^M n_m \text{cov}(X_h^{(m)} \underline{a}_h, Y_h^{(m)} b_h) + \lambda_h \|a_h\|_1,$$

a: vector of common loadings

MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms

Florian Rohart¹, Aida Estampé², Nicholas Matigian¹, Stéphanie Bougeard¹ and Kim-Anh Lê Cao^{1*}



In mgPLS, the PLS-components of each group are constraint to be built based on the same loading vectors in X and Y . These *global* loading vectors thus allow the samples from each group or study to be projected in the same common space spanned by the PLS-components.

We used a “Leave-One-Group-Out Cross-Validation (LOGOCV)”, which consists in performing CV where group or study m is left out only once $m = 1, \dots, M$. LOGOCV realistically reflects the true case scenario where prediction is performed on independent external studies based on a reproducible signature identified on the training set.

3 Vertical integration: in practice

```
R> library(mixOmics)
R> mint.pca(MyData,
  study = MyStudies)
R> mint.pls(MyData1, MyData2,
  study = MyStudies)
R> mint.plsda(MyData, OutCome,
  Study = MyStudies)
R> ...
```

- Case study:

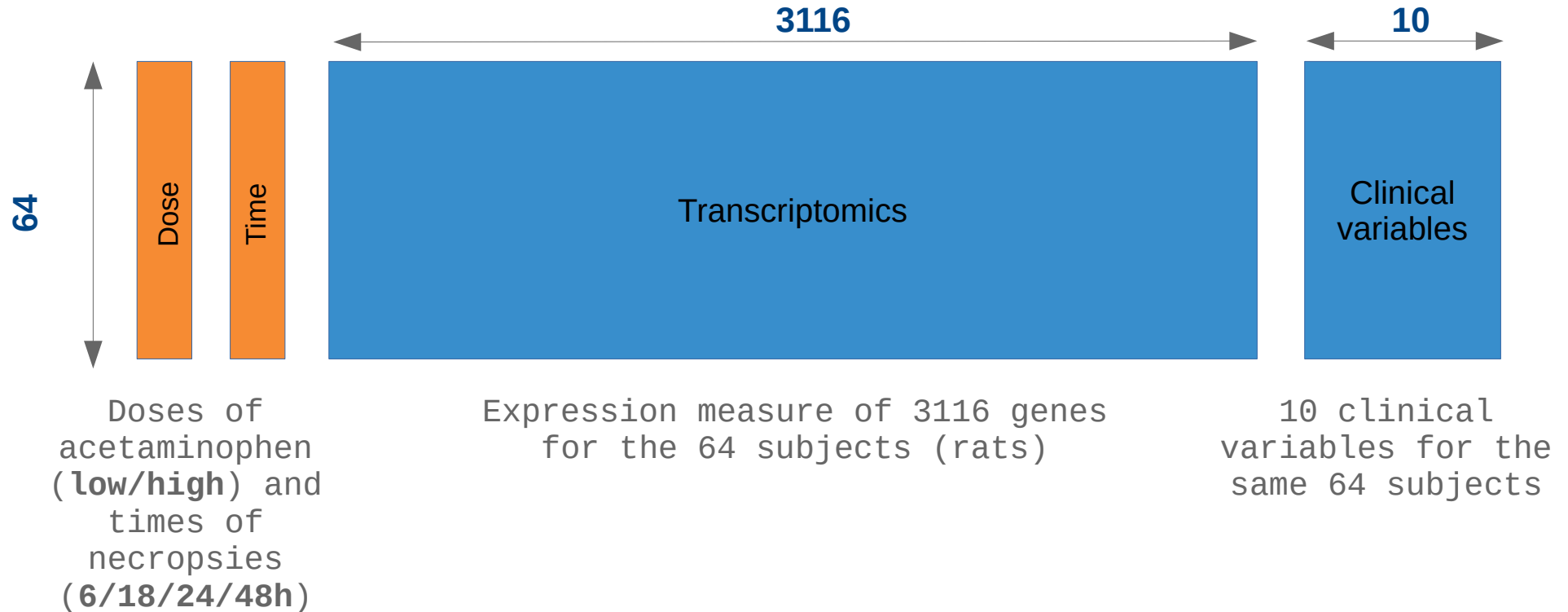
mixomics.org/mixmint/stemcells-example/



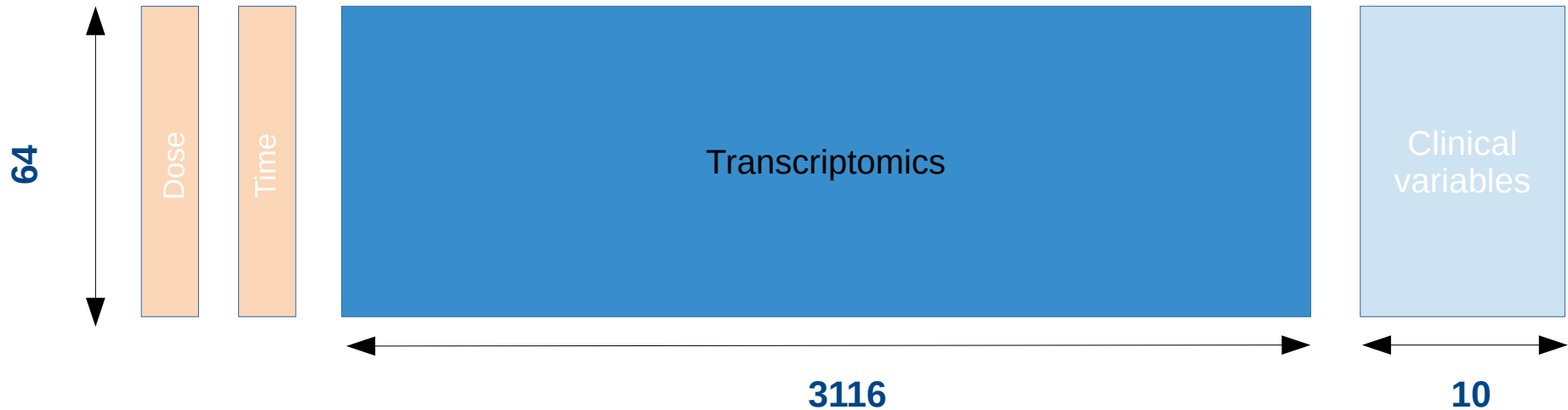
4 Example: Liver toxicity (LT)

```
R> library(mixOmics)
R> data(liver.toxicity)
R> help(liver.toxicity)
```

Bushel, P.R., Wolfinger, R.D. & Gibson, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Syst Biol* 1, 15 (2007). <https://doi.org/10.1186/1752-0509-1-15>

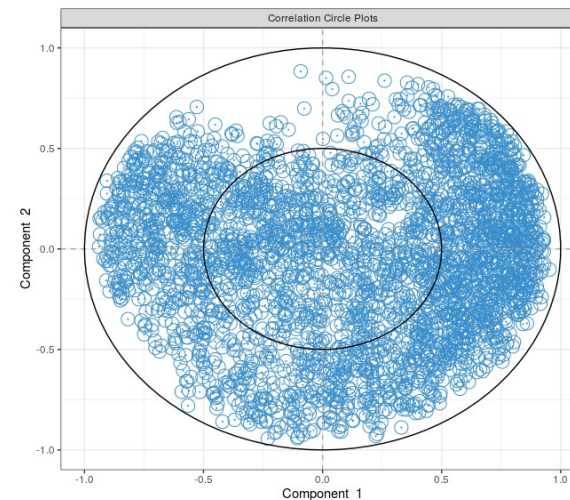
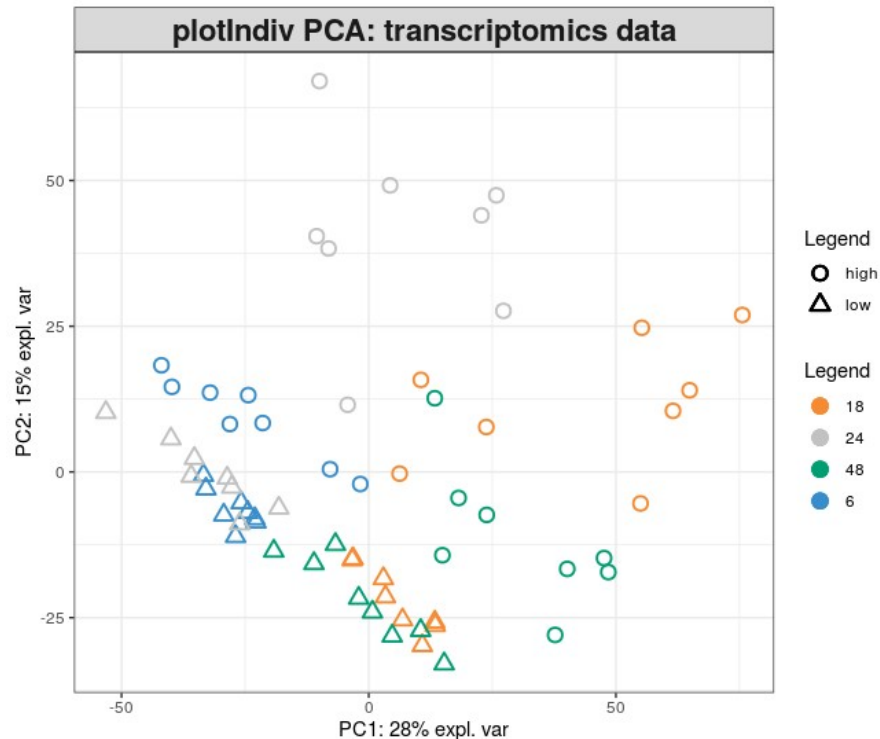


4 LT: explore one data set



Question: based on transcriptomics data, do we naturally observe clusters of samples which correspond to the different dose or exposure treatments?

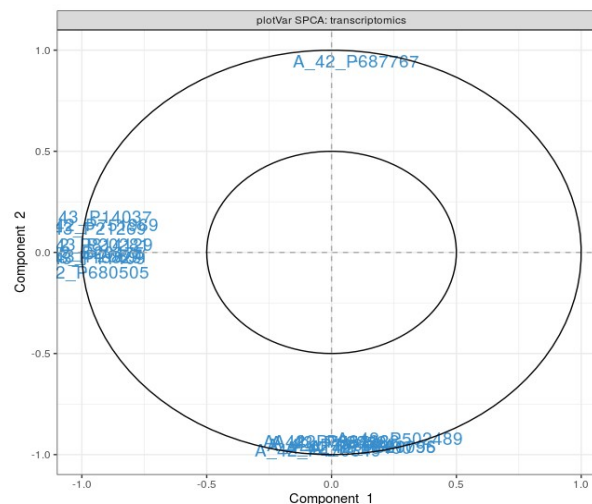
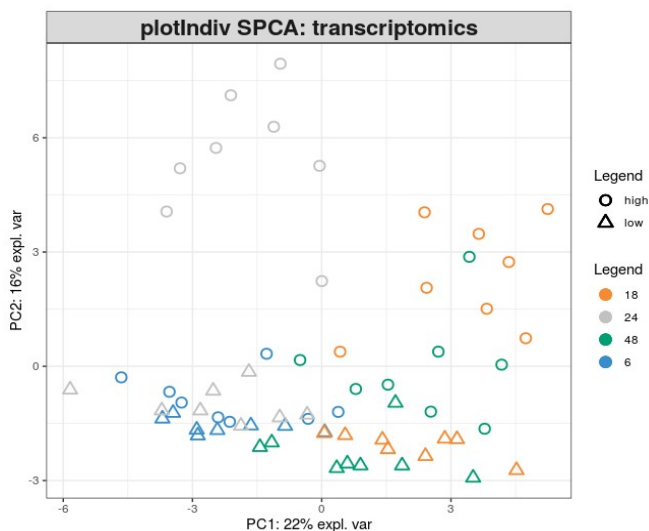
4 LT: PCA on transcriptomics data



Answer: dose effect appears clearly as well as trends in time effect...

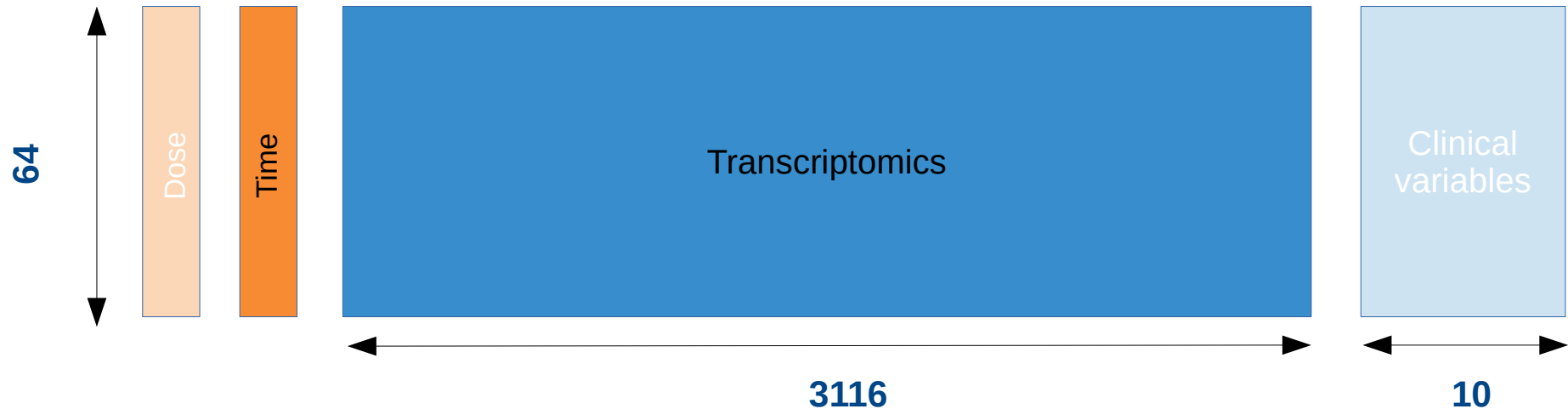
4 LT: Too many genes? Sparse PCA

Question: based on transcriptomics data, do we naturally observe clusters of samples which correspond to the different doses or exposure treatments **when we select some genes highly involved in the variability of the data?**



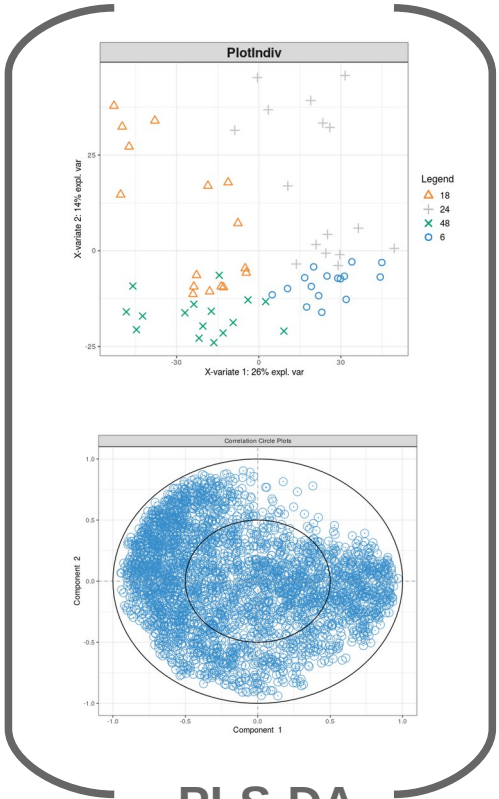
Answer: behaviour roughly similar when considering every gene or not.

4 LT: Supervised analysis: transcriptomics / time

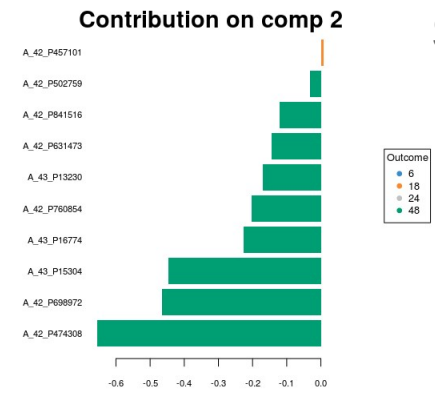
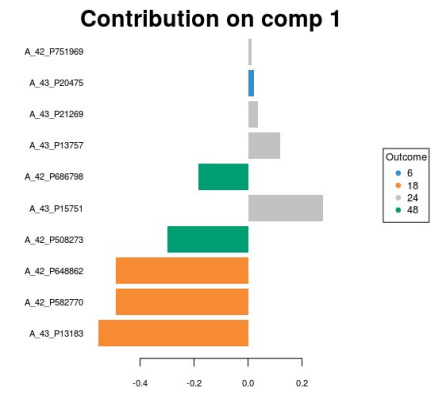


Question: Based on transcriptomics data, can we identify a molecular signature that characterizes the different treatment times?

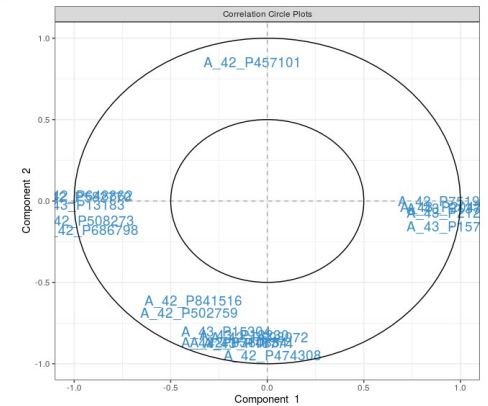
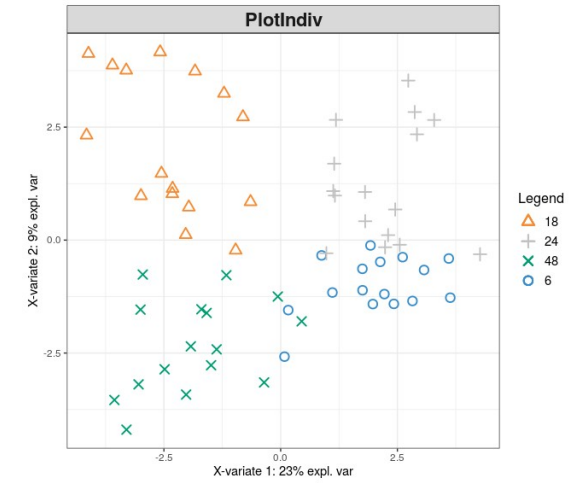
4 LT: (S)PLS-DA transcript. / time



PLS-DA

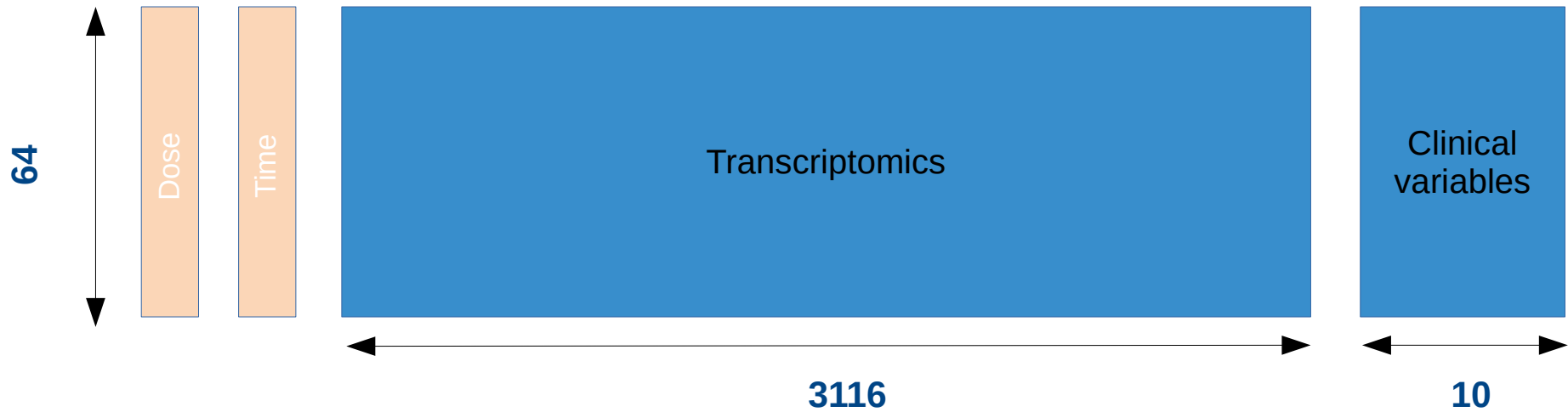


S-PLS-DA



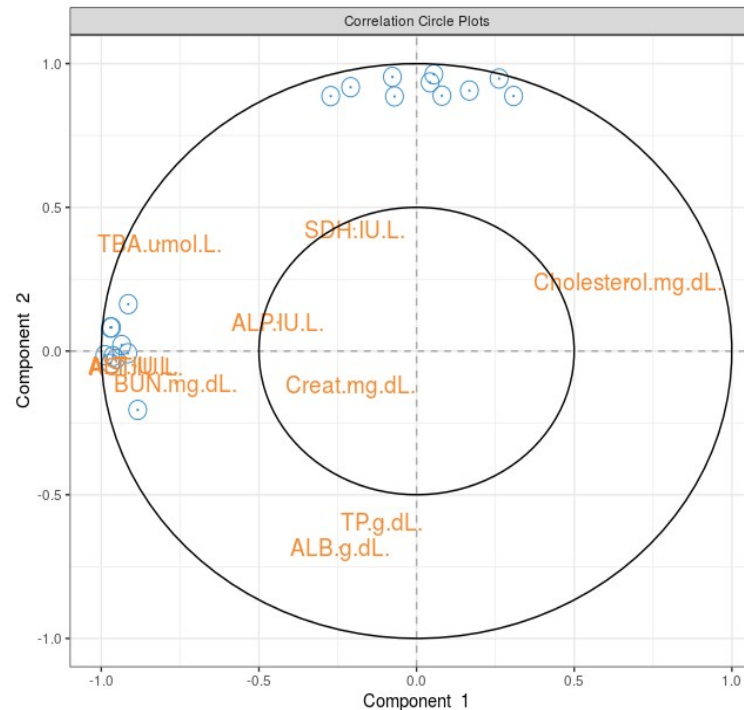
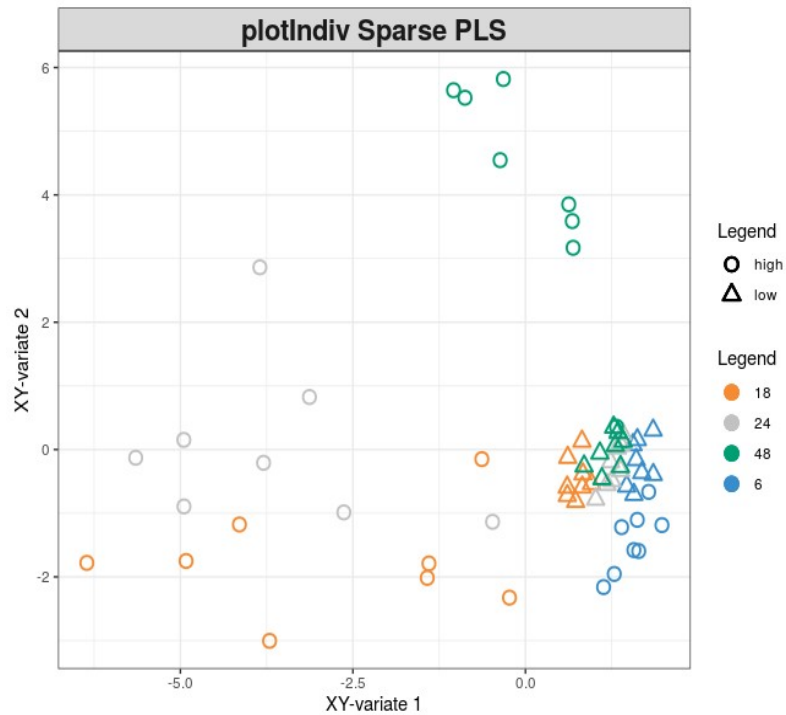
Answer: Probably, something to investigate...

4 LT: Unravel relationships between 2 datasets



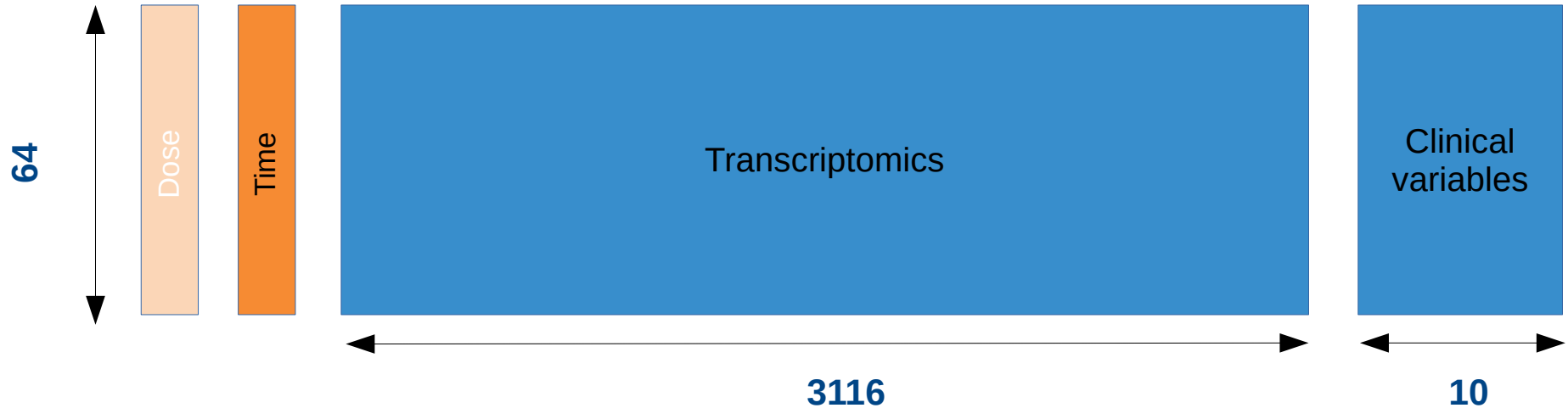
Question: Can we unravel relationships between transcriptomics data and clinical data ? **What are the genes that characterize these relationships?**

4 LT: Sparse PLS: transcriptomics / clinic



Answer: interesting trends on the individual plot and few genes involved.

4 LT: Multi-block supervised analysis



Question: Does the integration of the clinical and transcriptomics datasets bring better insight into the discrimination of the samples based on the time of necropsies?

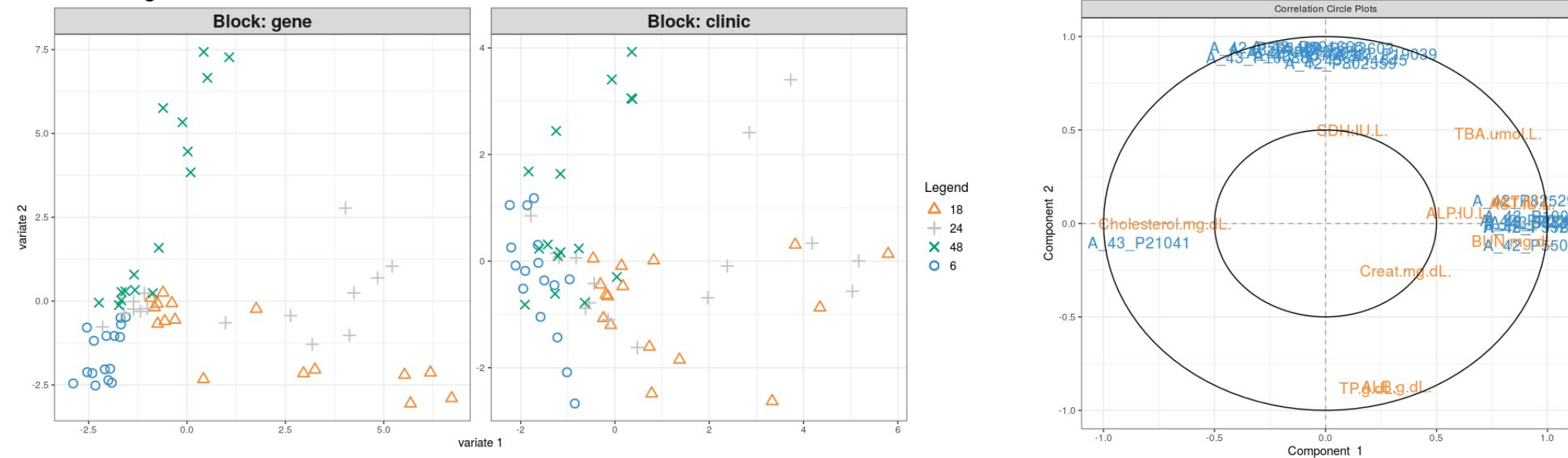
Investigation carried out with two design matrices

	Full design		
	Tr.	Cl.	Time
Trans.	0	1	1
Clinic.	1	0	1
Time	1	1	0

	DA-oriented design		
	Tr.	Cl.	Time
Trans.	0	0.1	1
Clinic.	0.1	0	1
Time	1	1	0

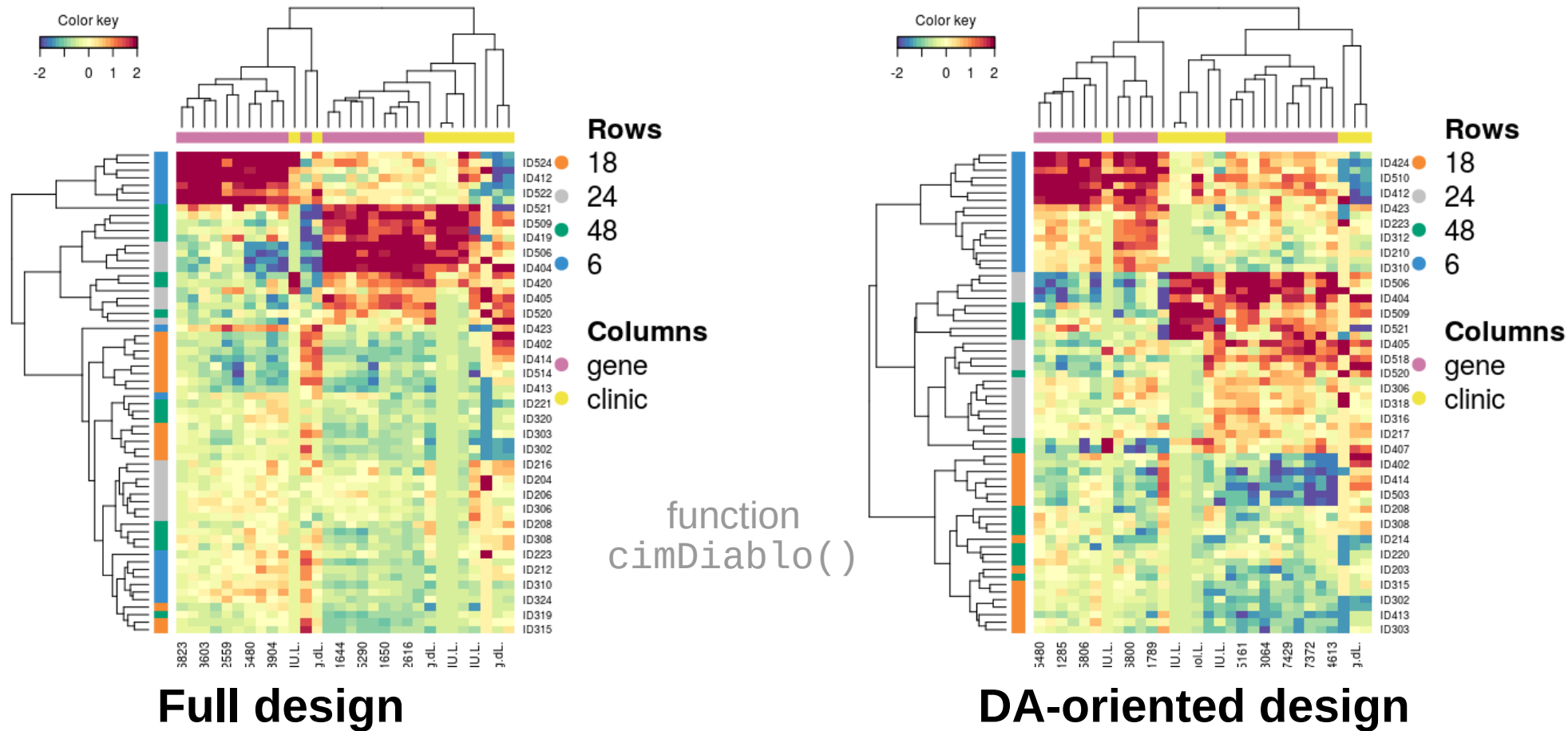
4 LT: Multi-block sparse PLS-DA: transcriptomics / clinic / time

Full design

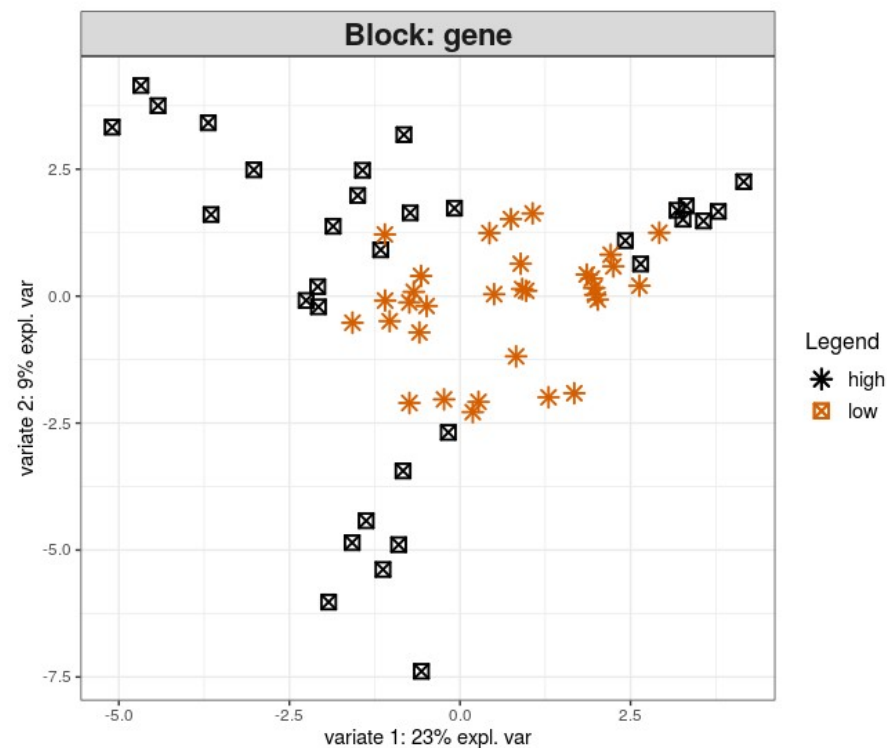
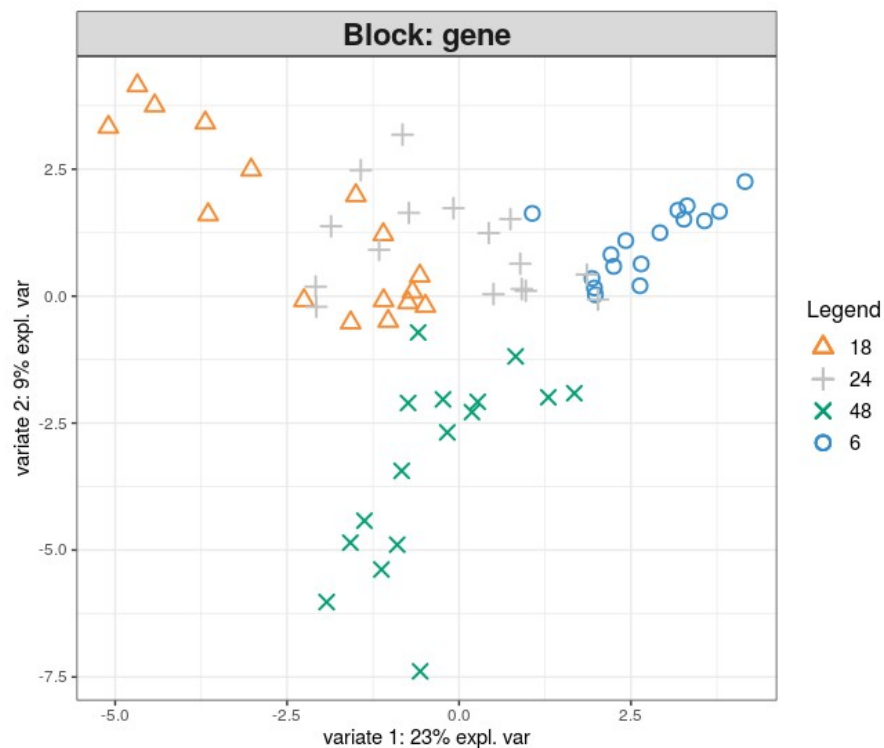


Answer: results to be investigated...

4 LT: Multi-block sparse PLS-DA: transcriptomics / clinic / time



4 LT: Multi-block sparse PLS-DA: transcriptomics / clinic / time



DA-oriented design

4 Example: Wallomics



Laboratoire de Recherche en Sciences Végétales

www.lrsv.ups-tlse.fr

- **60** samples *A. thaliana*:
 - **5** ecotypes (Col, Hosp, Grip, Hern, Roch)
 - **2** temperatures (low, high)
 - **2** organs (stem, rosette)
 - **3** replicates
- **4** data sets: proteomics (400), transcriptomics (20000), metabolomics-sugar (7), phenomics (9)



Take home message

- Practice on your own data! The best way to understand what a method has to tell you.
- Do not bypass the elementary analyses (univariate, bivariate, multivariate single data set).
- Address problems explicitly formulated: “I want to integrate my data” is not a problem explicitly formulated.
- Clearly identify supervised and unsupervised questions and the methods to use.