Multivariate projection methodologies for the exploration of large biological data sets

Examples

mixOmics

Exploration and

Integration of

Omics datasets

Genotoul Biostat

math.univ-toulouse.fr/biostat

Agenda

- Introduction

- **Nutrimouse** data set: PCA, (Sparse-)PLS
- **SRBCT** data set: (Sparse-)PLS-DA
- **WallOmics** data set: multiblock-(sparse-)PLS, DIABLO

- Conclusion

Introduction

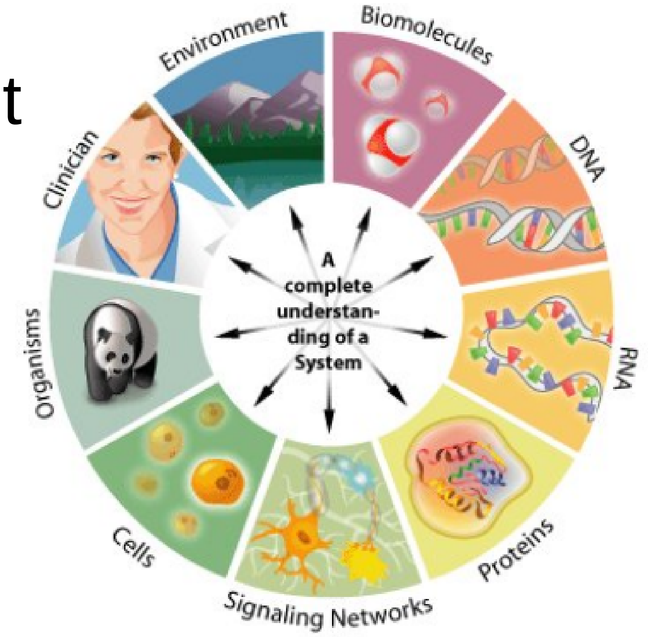
Research context

- From reductionism:

  1 gene = 1 hypothesis = 1 statistical test

- To holism:

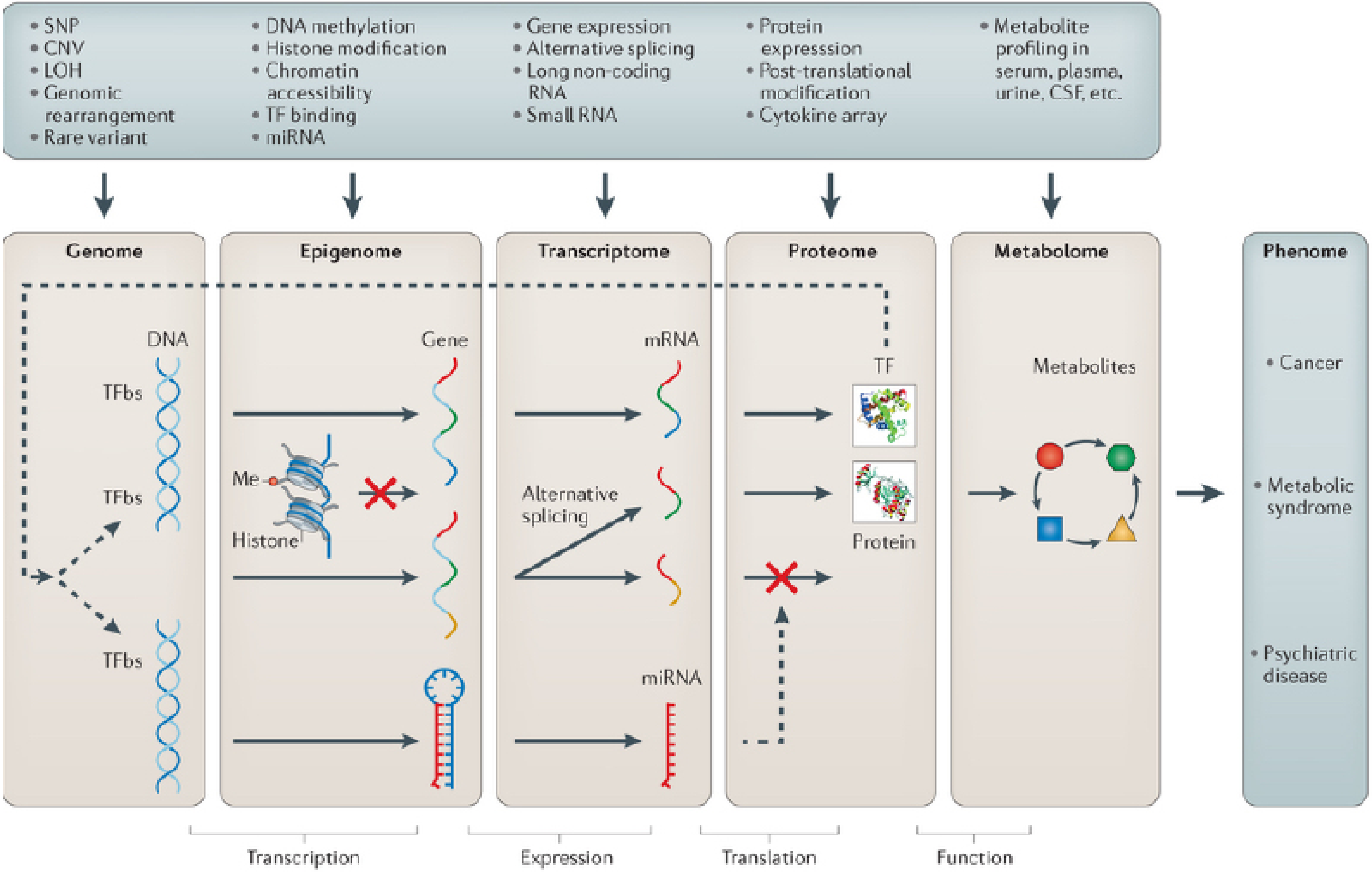
  Thousands of molecules = ??



- Biological aims:
  - integrate data from different ‘omics molecular levels to better understand a biological system
  - postulate novel biological hypotheses to be validated in the lab

Heterogeneous omics data

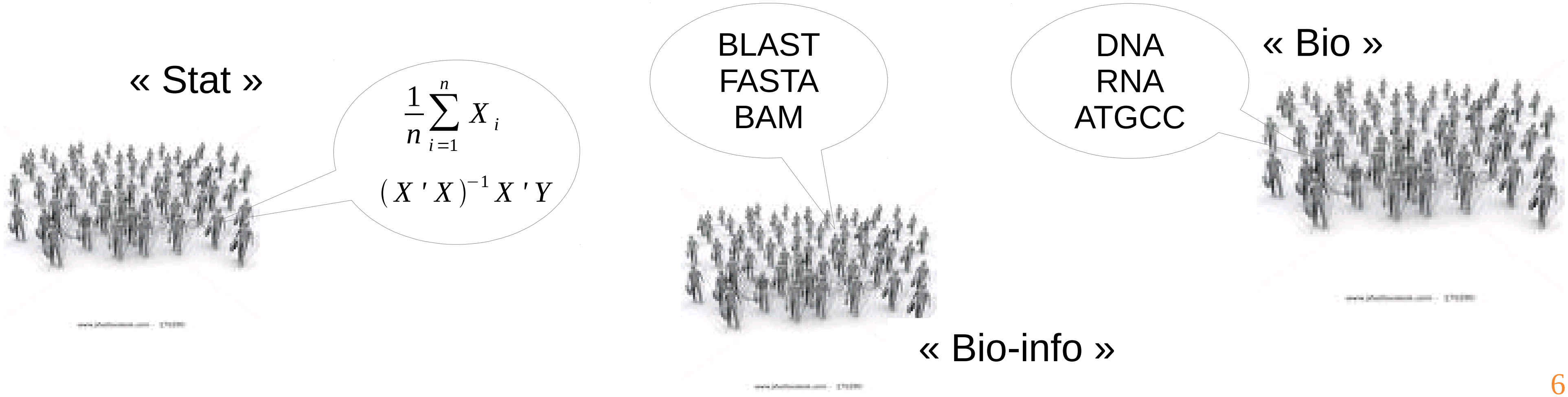
From genome to phenome



Multidisciplinarity!

- Nearly unlimited quantity of data from multiple and heterogeneous sources
- Computational issues to foresee
- Biological interpretation for validation
- Keep pace with new technologies

A close interaction between statisticians, bioinformaticians and molecular biologists is essential to provide meaningful results



Research hypothesis

- Molecular entities act together to trigger cells’ responses and need to be appropriately modelled and identified using novel statistical techniques.

- Multivariate statistical methods to shift the univariate statistics paradigm to obtain deeper insight into biological systems
  - Identify a combination of biomarkers rather than univariate biomarkers
  - Integrate multiple sources of biological data
  - Reduce the dimension of the data for a better understanding of complex biological systems

Data integration

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)

mixOmics philisophy in this context:

- R toolkit for multivariate data analysis of ‘omics data
- Statistical data integration
- Data-driven approaches (≠ database or knowledge-based approaches)

Nutrimouse

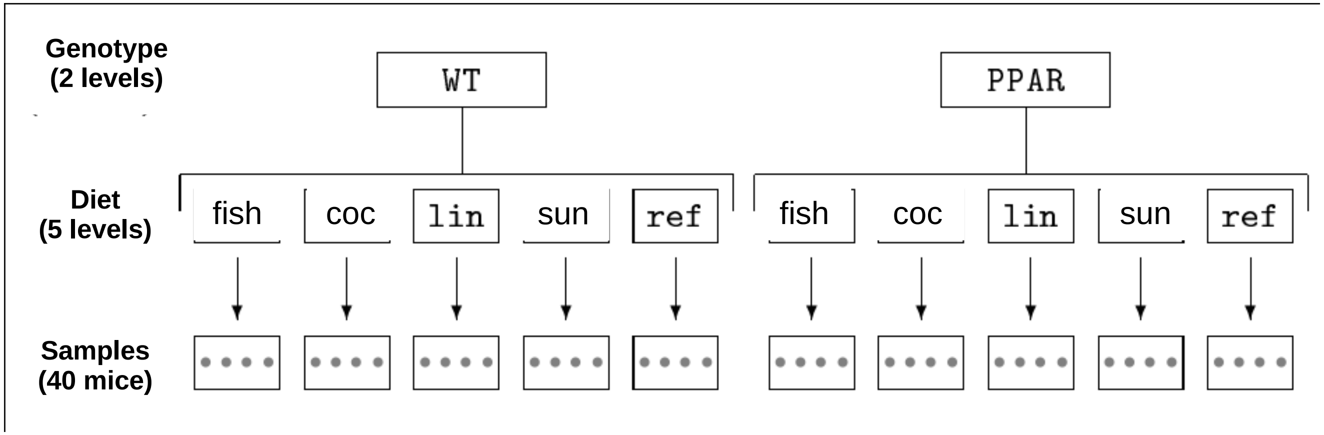
Experimental design

RESEARCH CENTRE IN FOOD TOXICOLOGY

Alimentation Humaine
&
Santé Animale

Pascal Martin, Thierry Pineau, Inra ToxAlim

- **40** mice: **2** genotypes x **5** diets x **4** replicates



Ignacio González

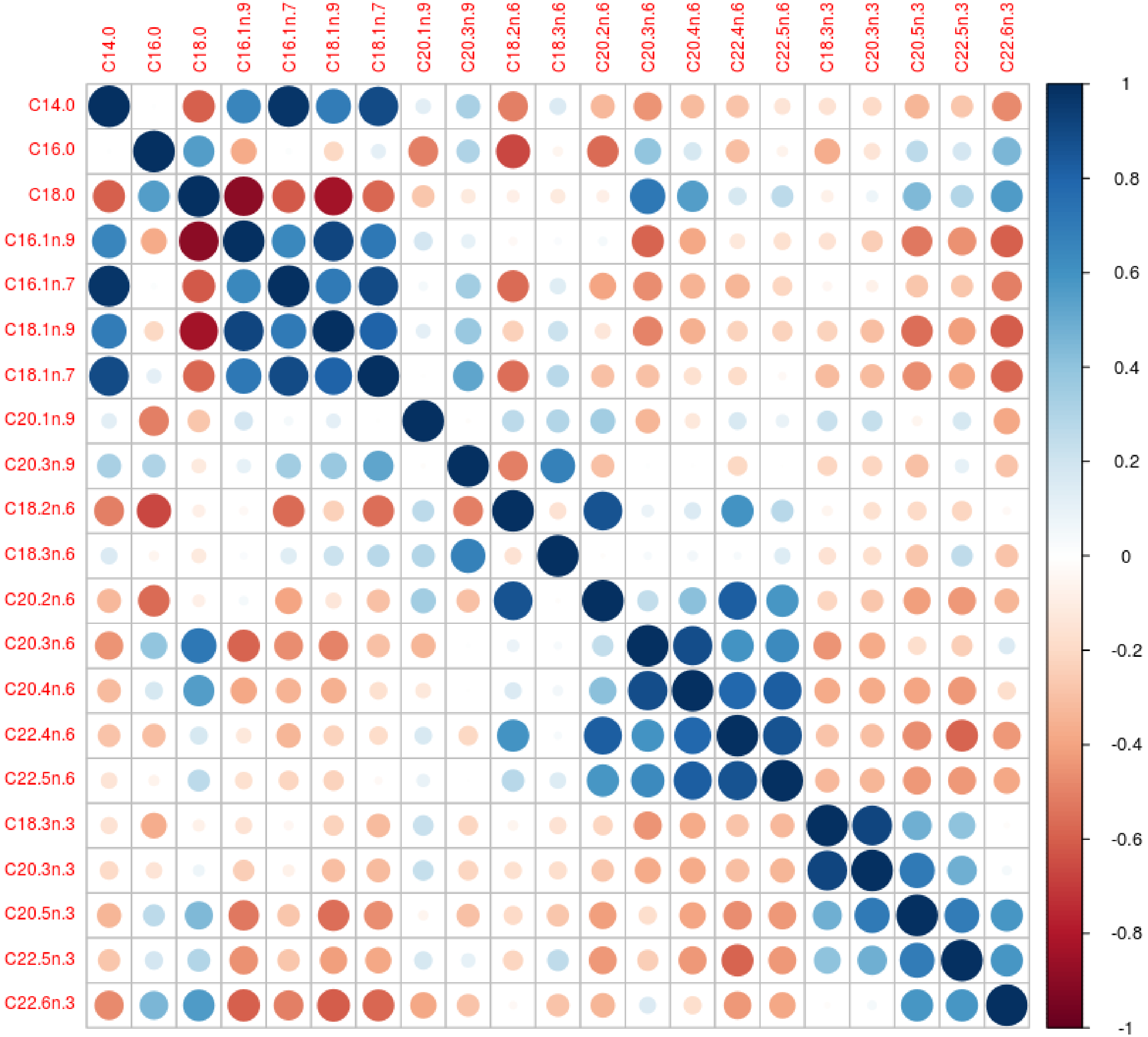
Oils used for experimental diets preparation were corn and colza oils (50/50) for a reference diet (**REF**), hydrogenated coconut oil for a saturated fatty acid diet (**COC**), sunflower oil for an Omega6 fatty acid-rich diet (**SUN**), linseed oil for an Omega3-rich diet (**LIN**) and corn/colza/enriched fish oils for the **FISH** diet (43/43/14).

- **2** data sets: **21** hepatic fatty acids and expression of **120** genes in liver cells

Martin, P. G. P., Guillou, H., Lasserre, F., Dejean, S., Lan, A., Pascussi, J.-M., San Cristobal, M., Legrand, P., Besse, P. and Pineau, T. (2007). Novel aspects of PPARÎ±-mediated regulation of lipid and xenobiotic metabolism revealed through a multrigenomic study. *Hepatology*, 54, 767-777.

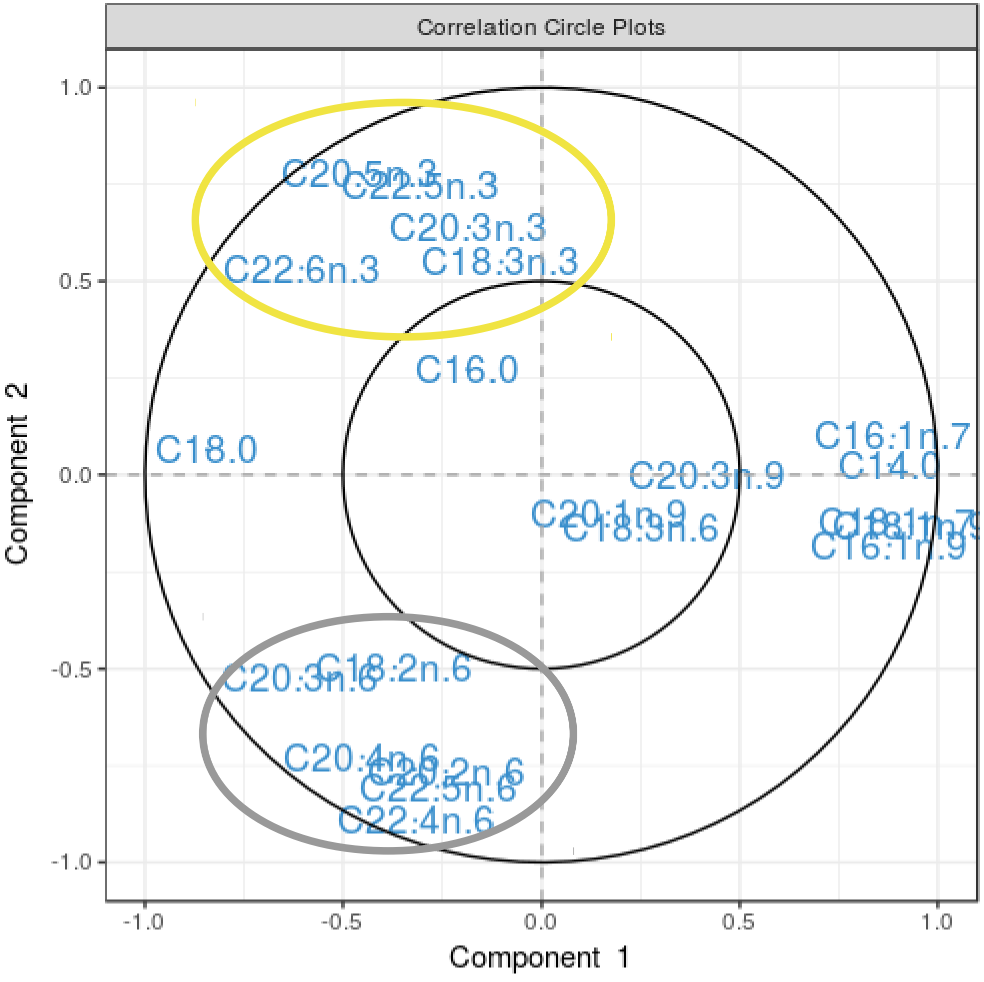
lipids	genes

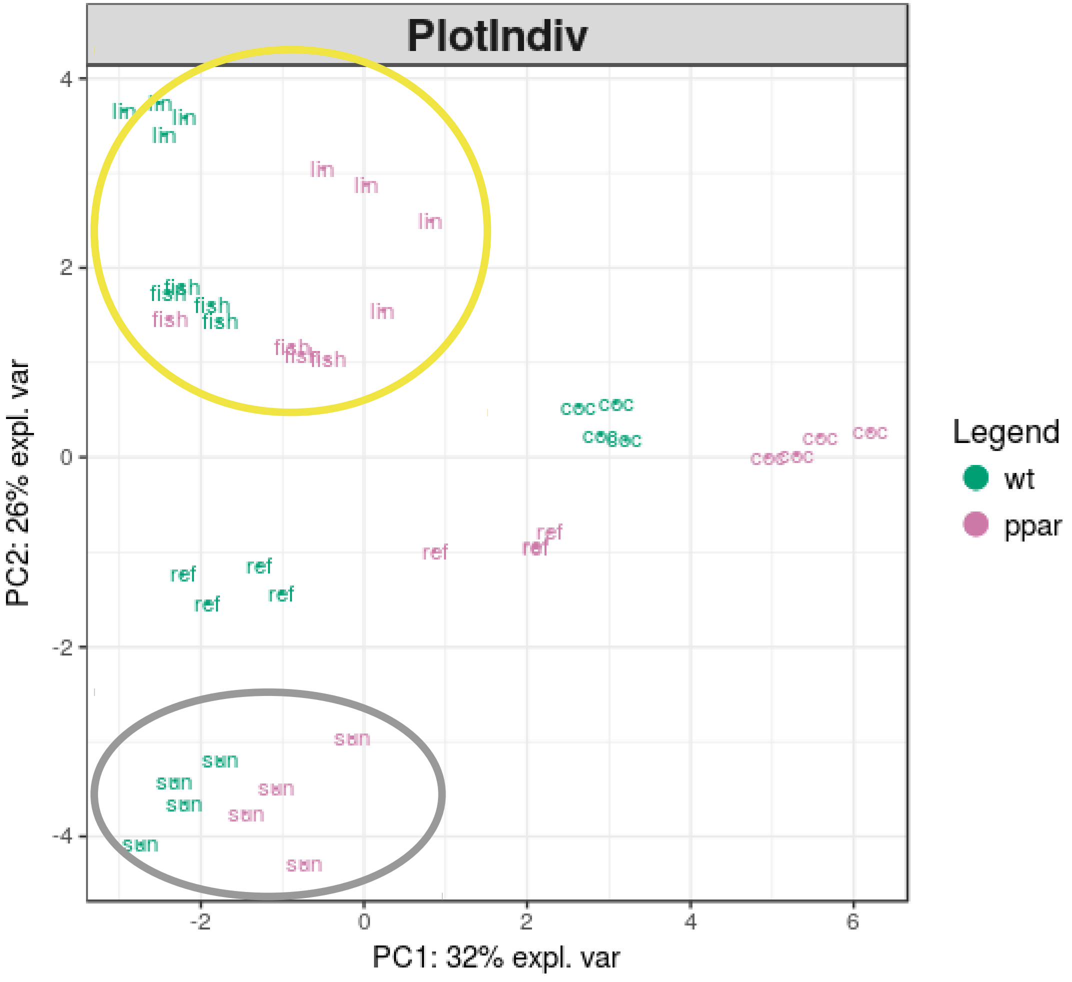
Correlations: lipids



Package corrplot

PCA lipids



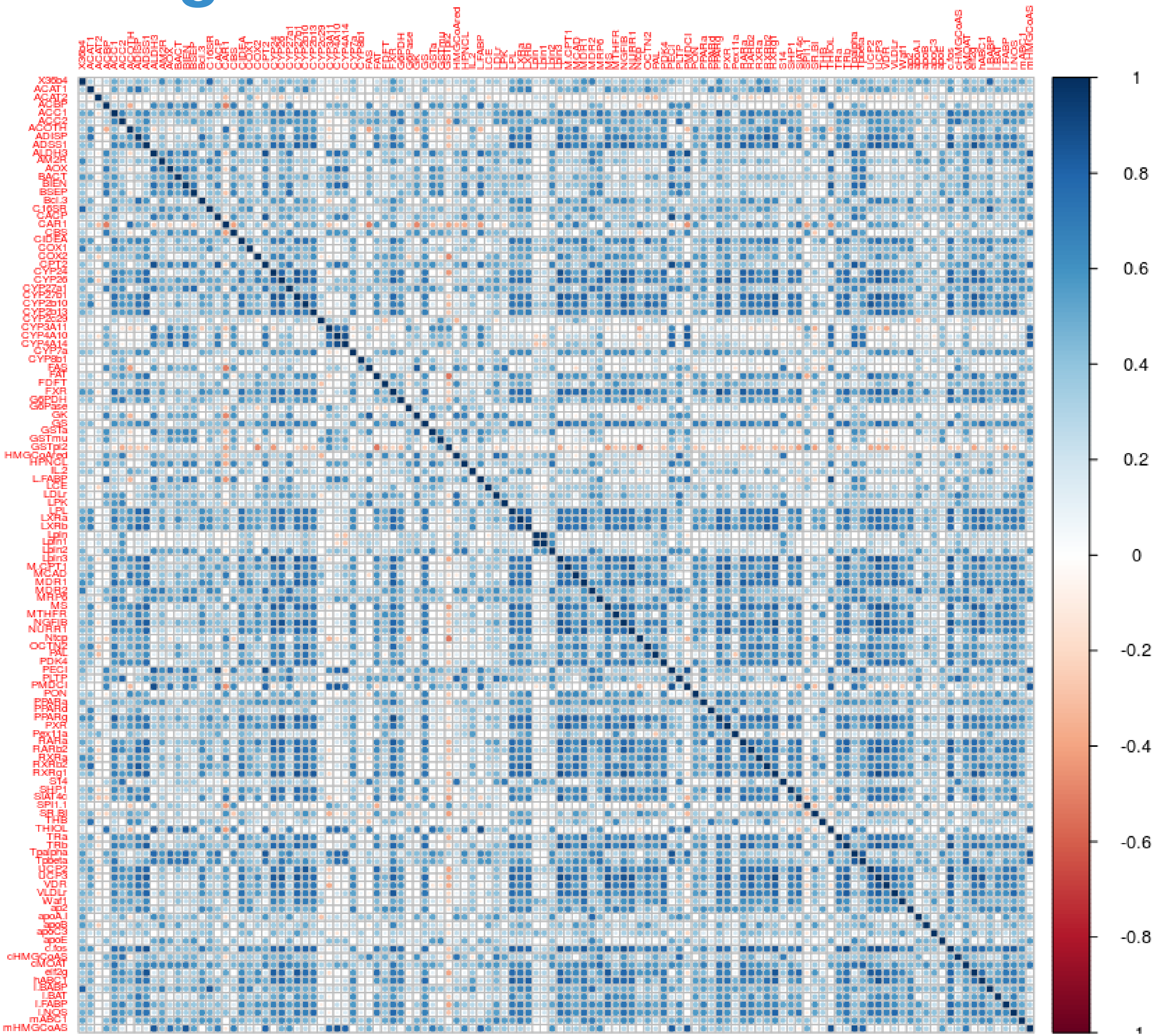


Variables plot

Individuals plot

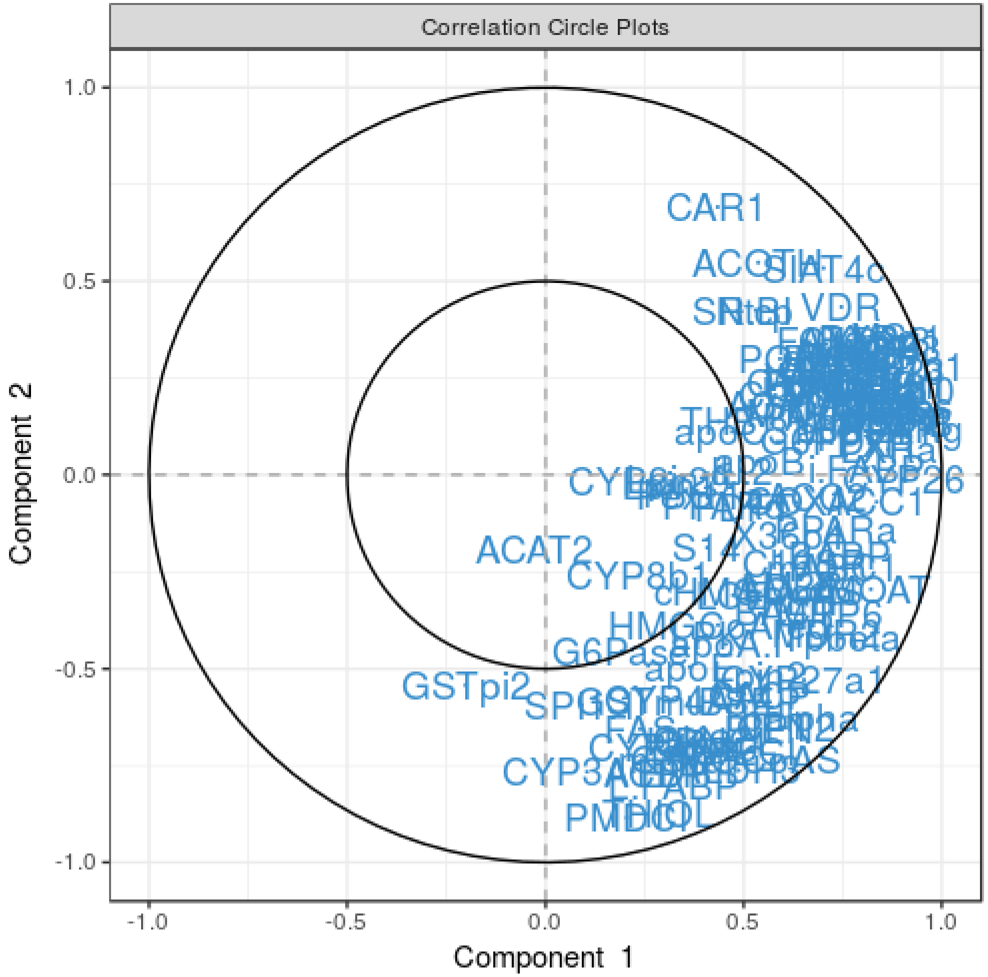
Introduction | Nutrimouse | SRBCT | WallOmics | Conclusion

Correlations: genes

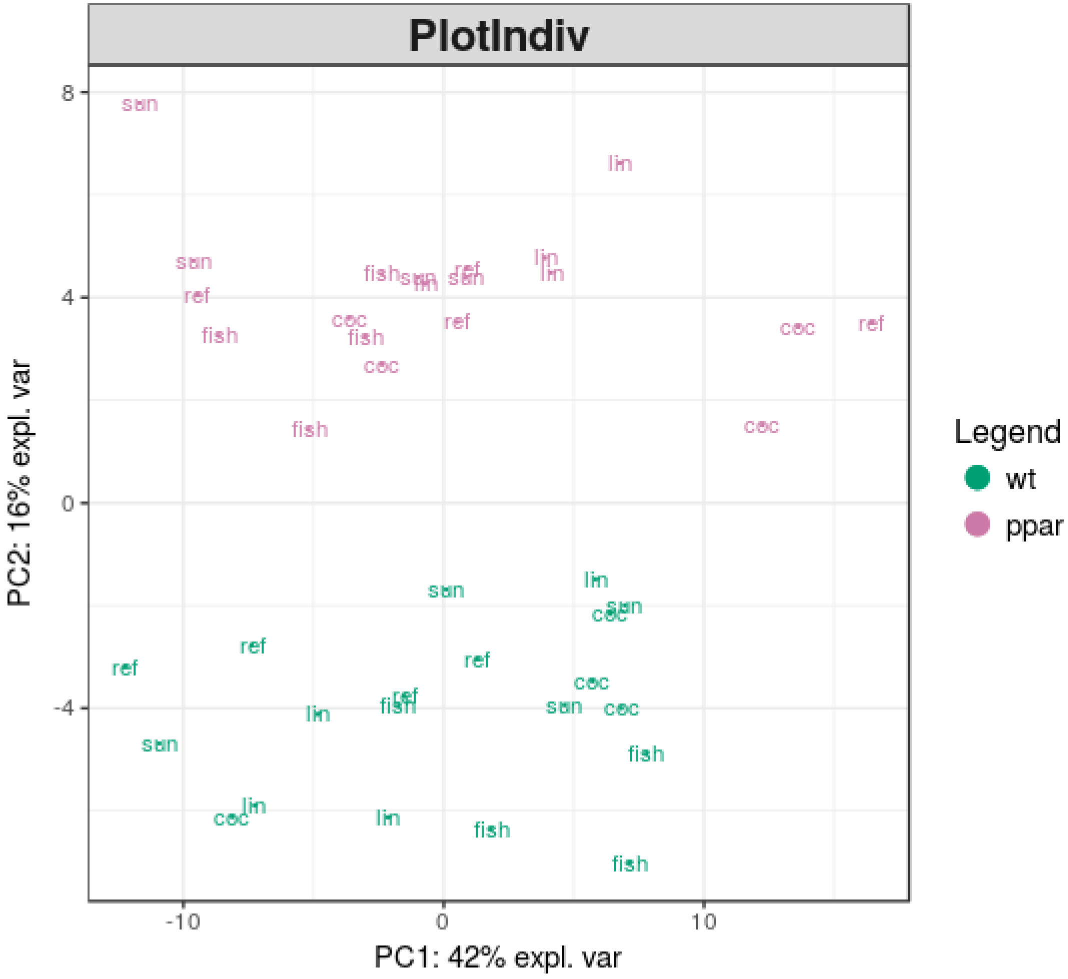


Package
corrplot

PCA genes

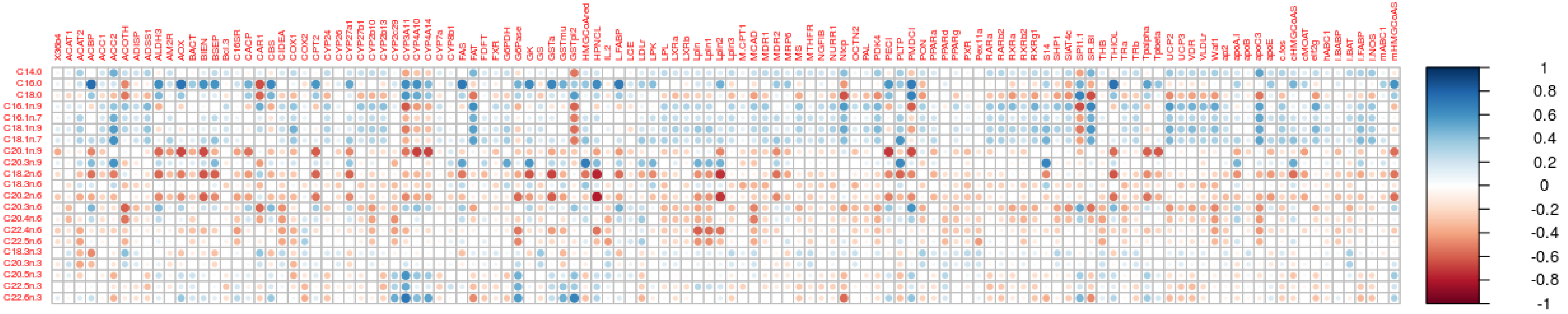


Variables plot



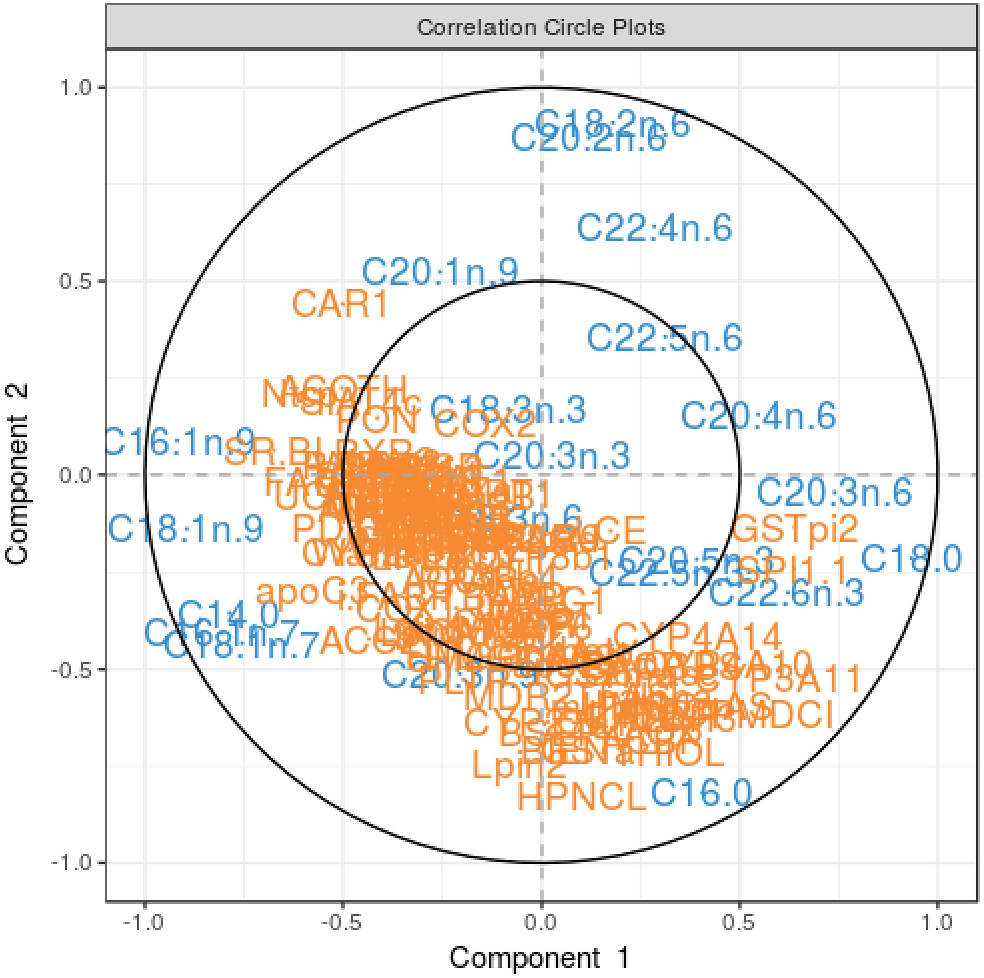
Individuals plot

Correlations: genes and lipids

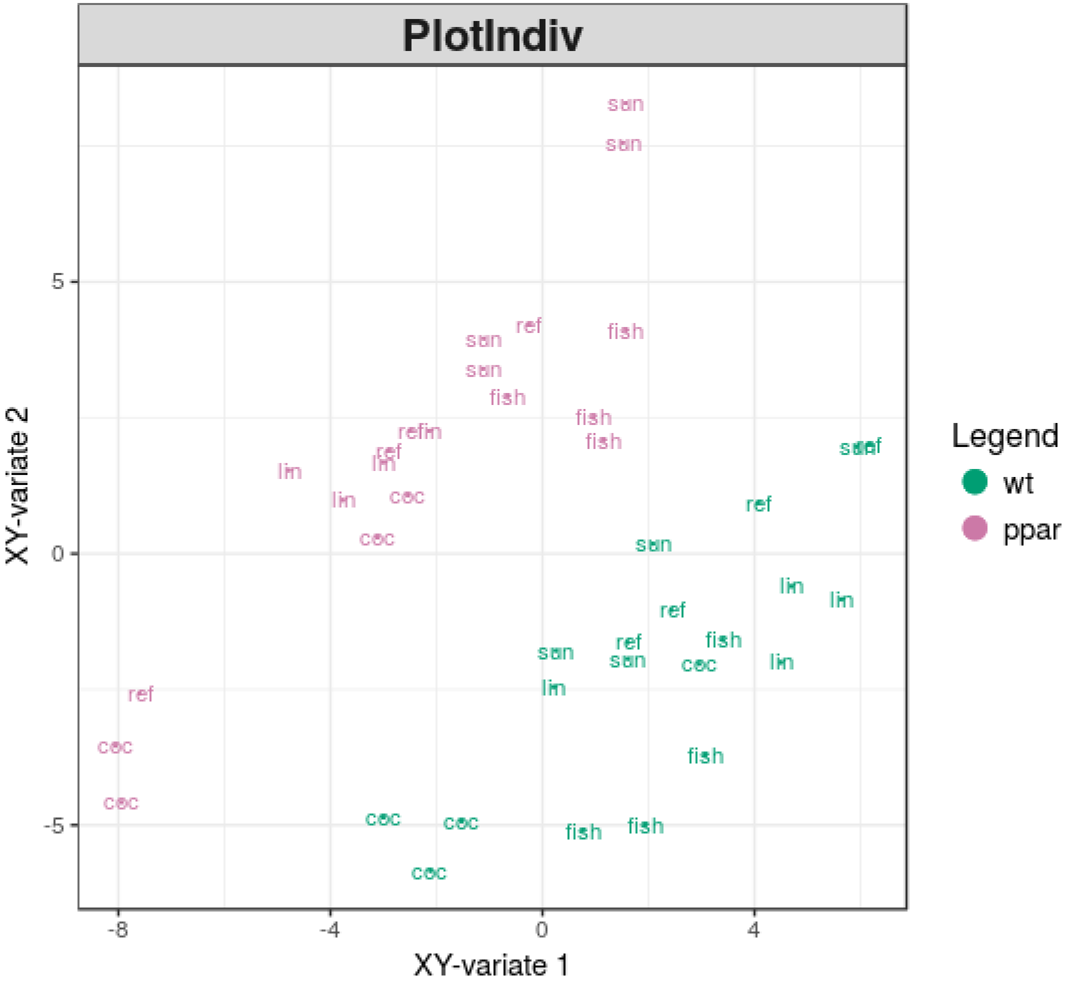


Package `corrplot`

PLS

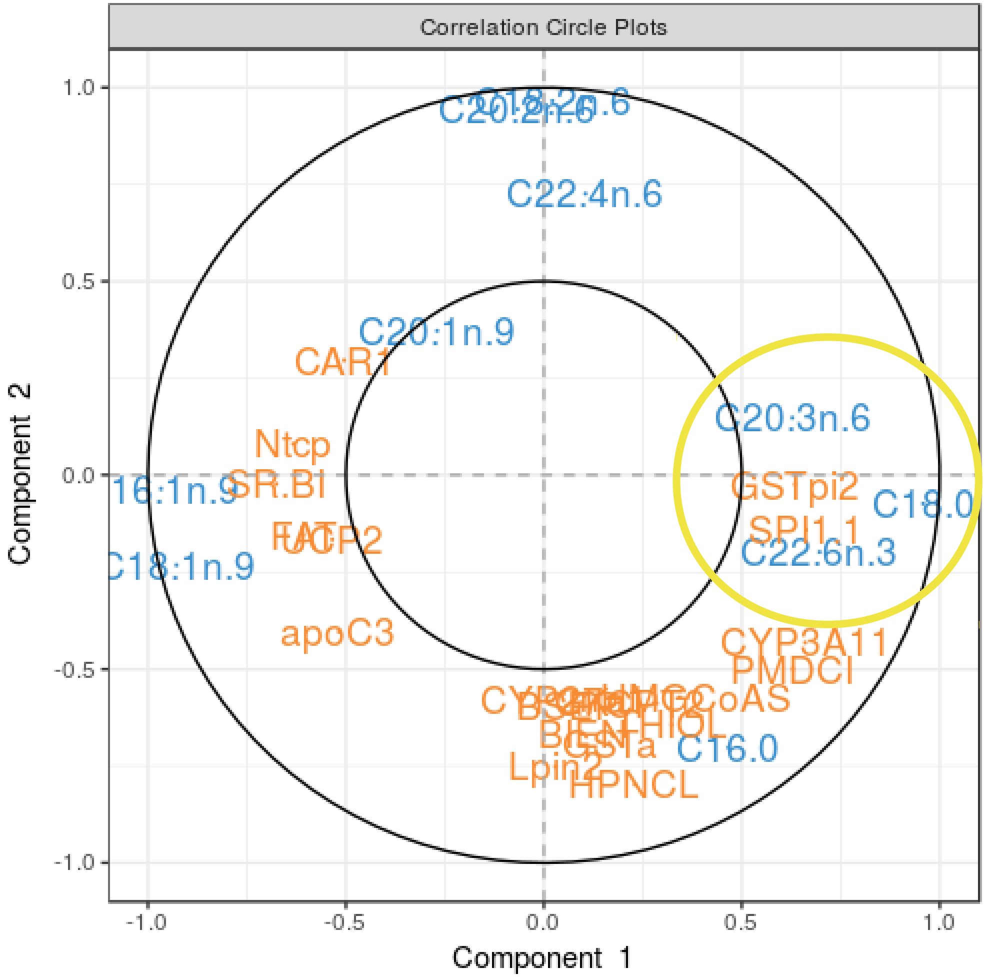


Variables plot

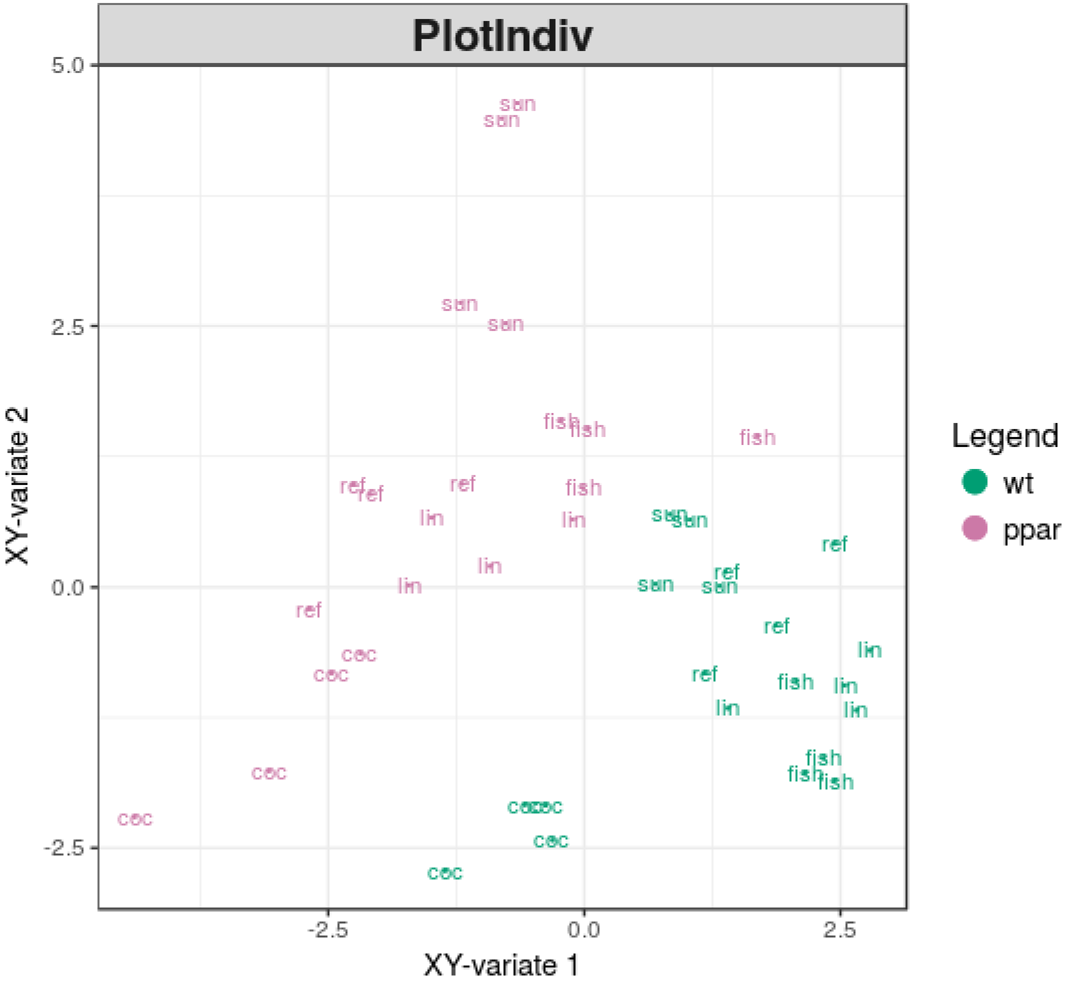


Individuals plot

Sparse PLS

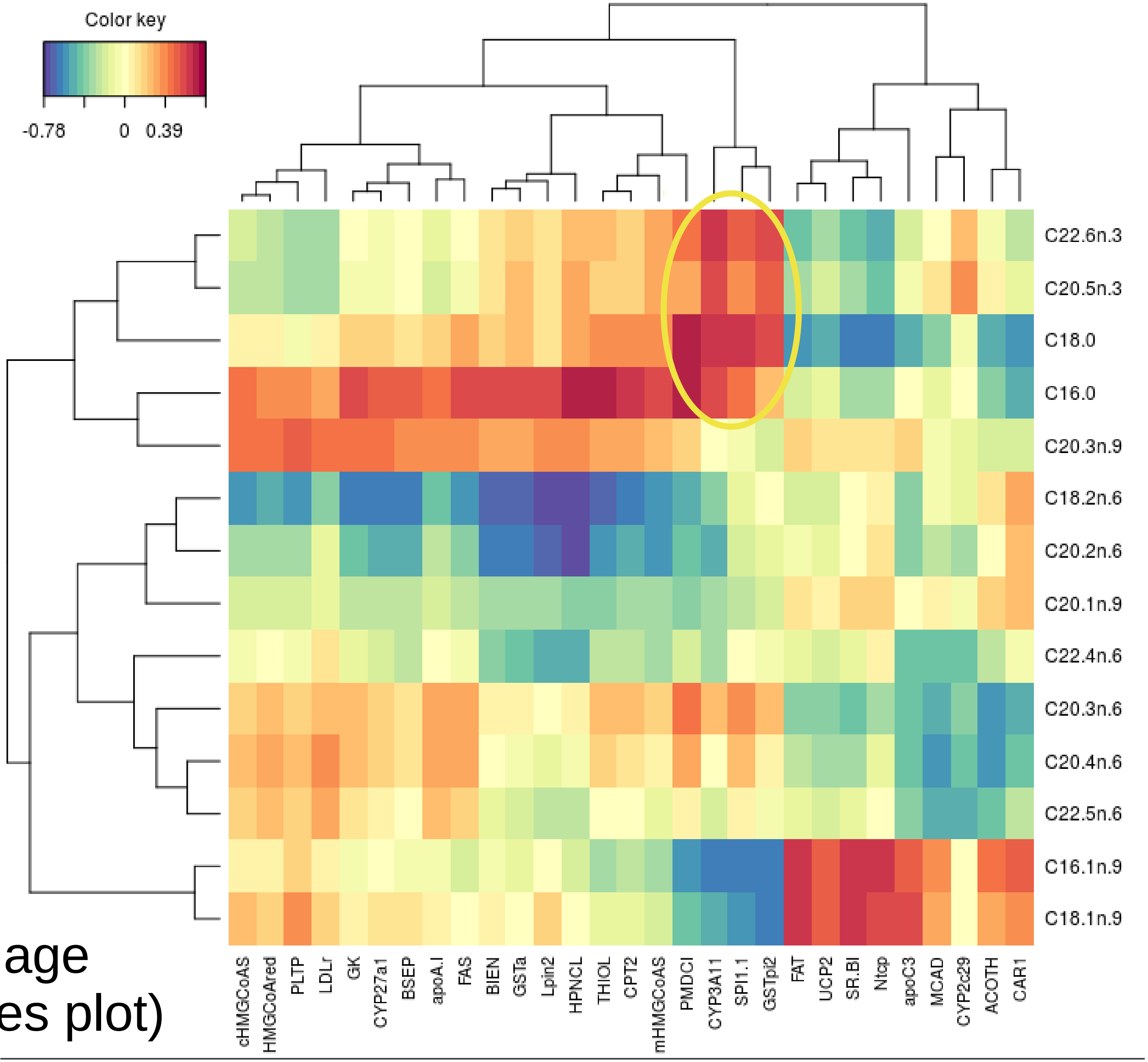


Variables plot



Individuals plot

Sparse PLS



Clustered Image Map (variables plot)

Small Round Blue Cell Tumors (SRBCT)

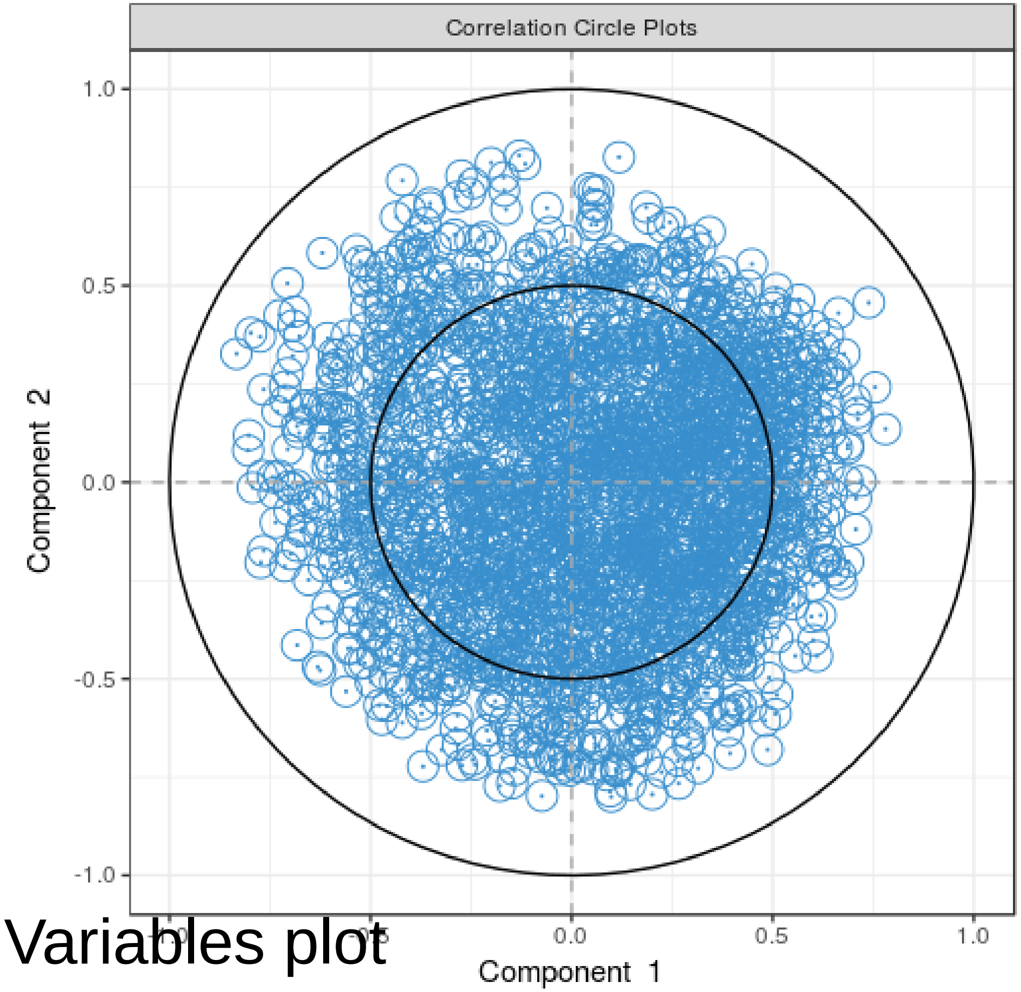
Experimental design

- **63** subjects
- Expression of **2308** genes
- Class tumour of each subject, 4 classes: 23 Ewing Sarcoma (EWS), 8 Burkitt Lymphoma (BL), 12 neuroblastoma (NB), 20 rhabdomyosarcoma (RMS)

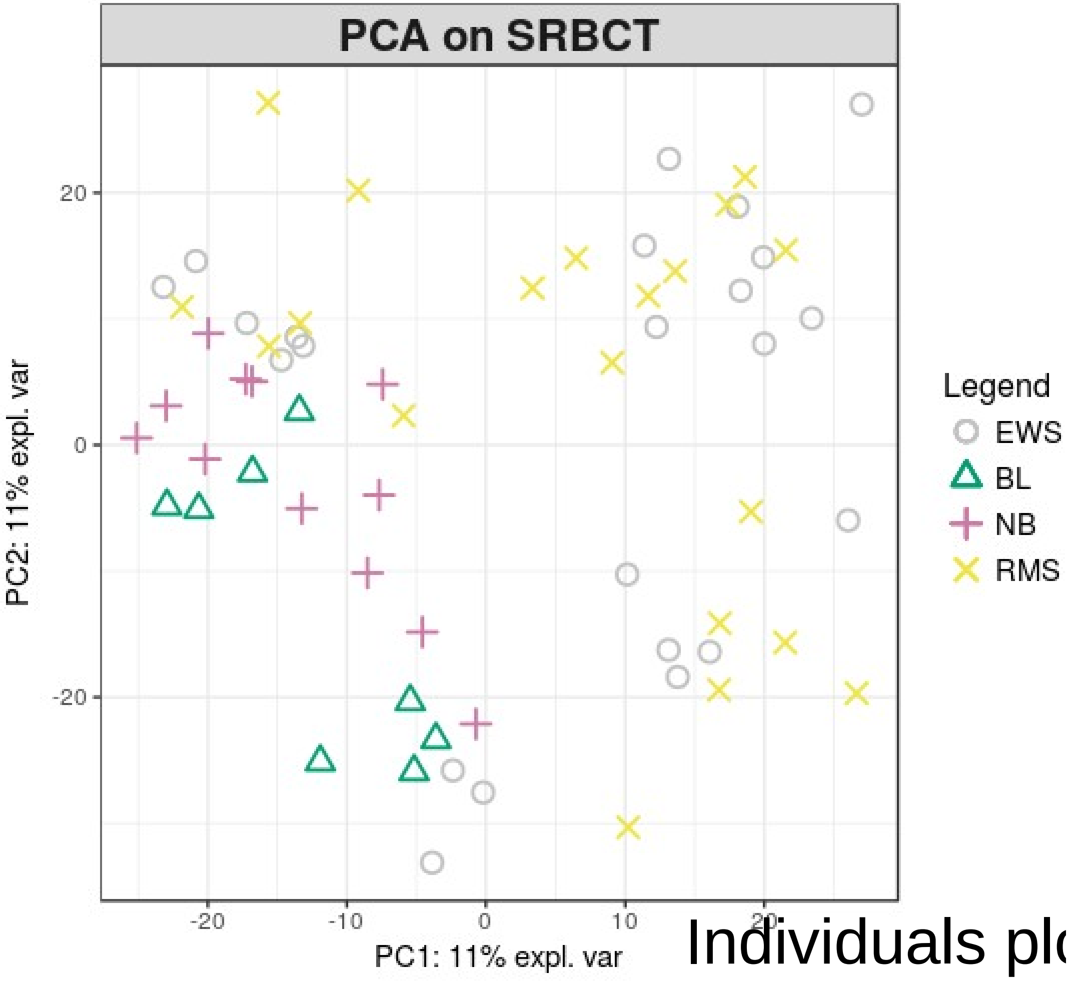
Khan et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.*Nature Medicine* 7, Number 6, June.

https://research.nhgri.nih.gov/microarray/Supplement/

PCA

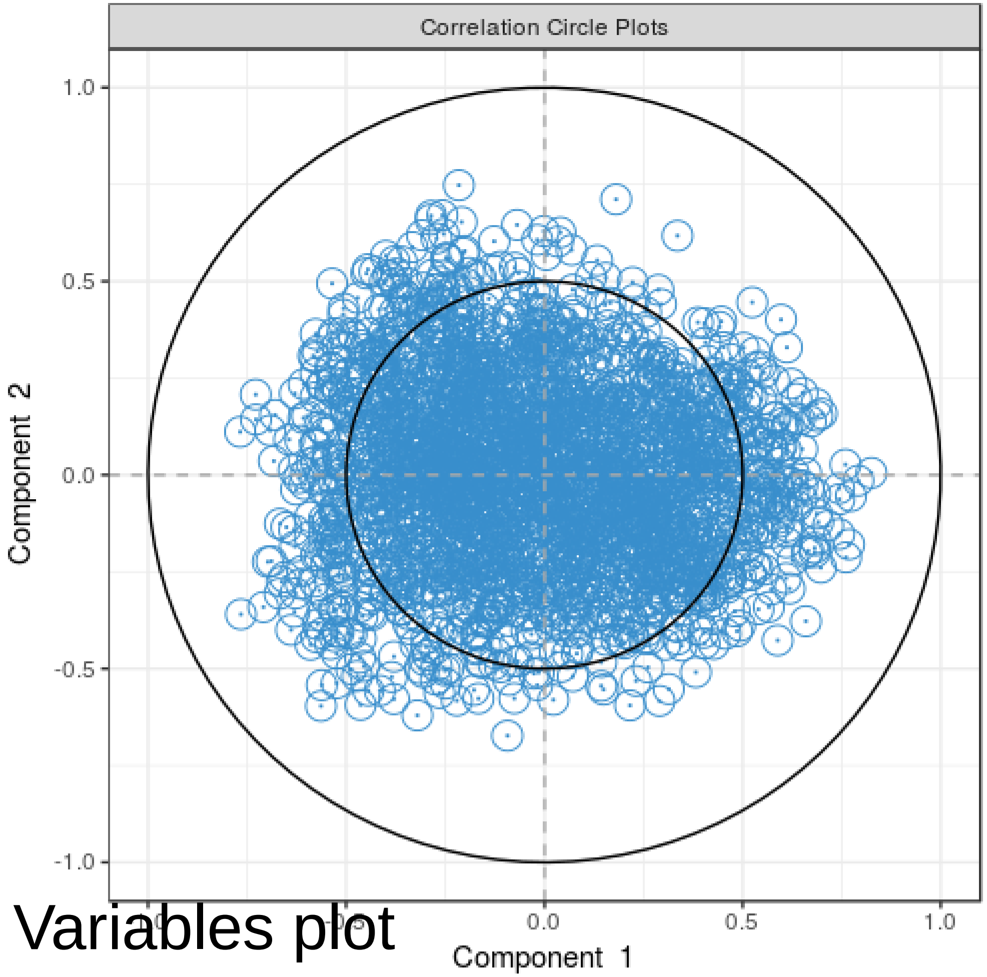


Variables plot

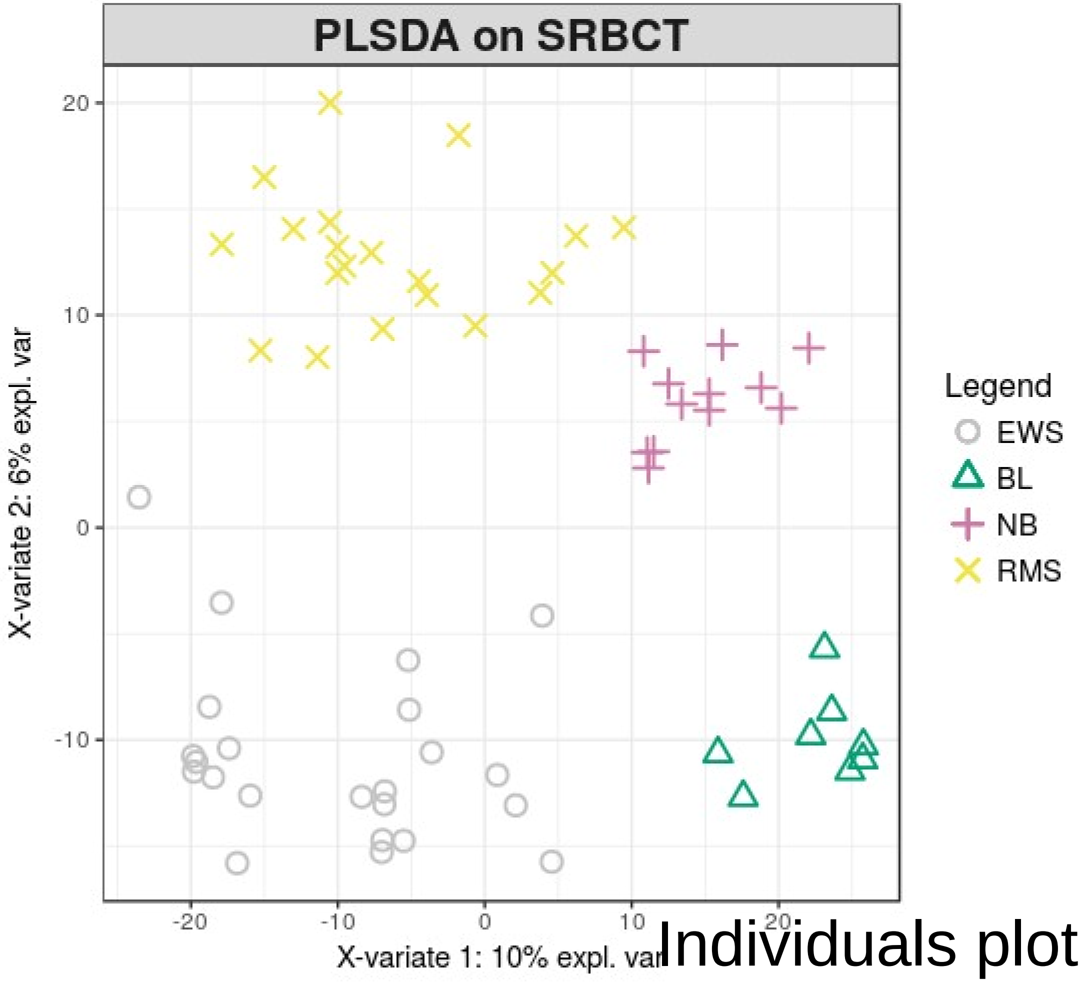


Individuals plo

PLS-DA

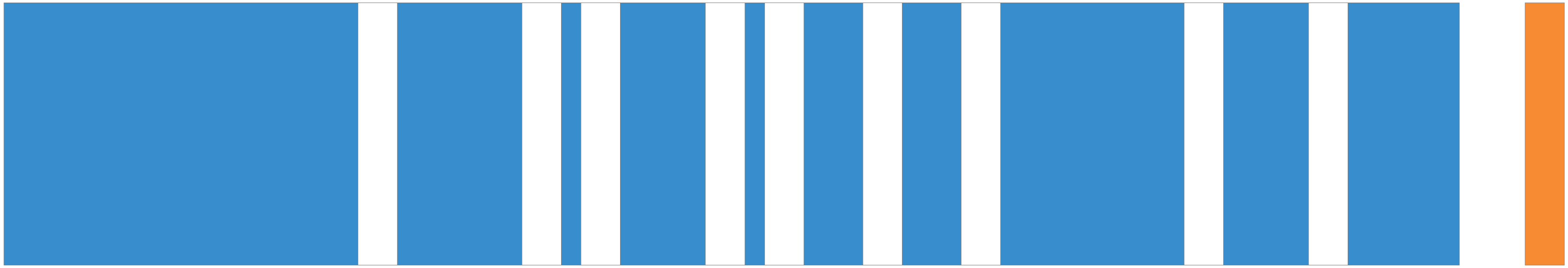


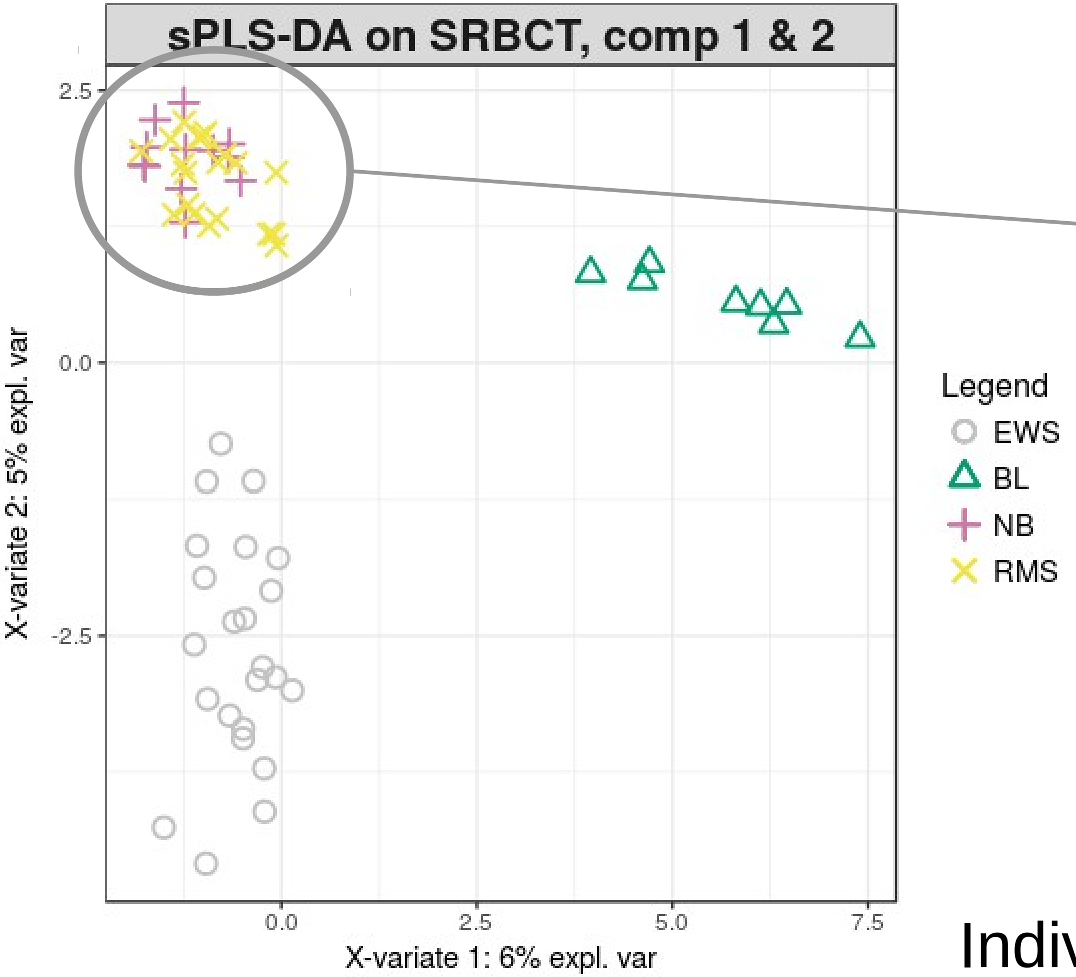
Variables plot

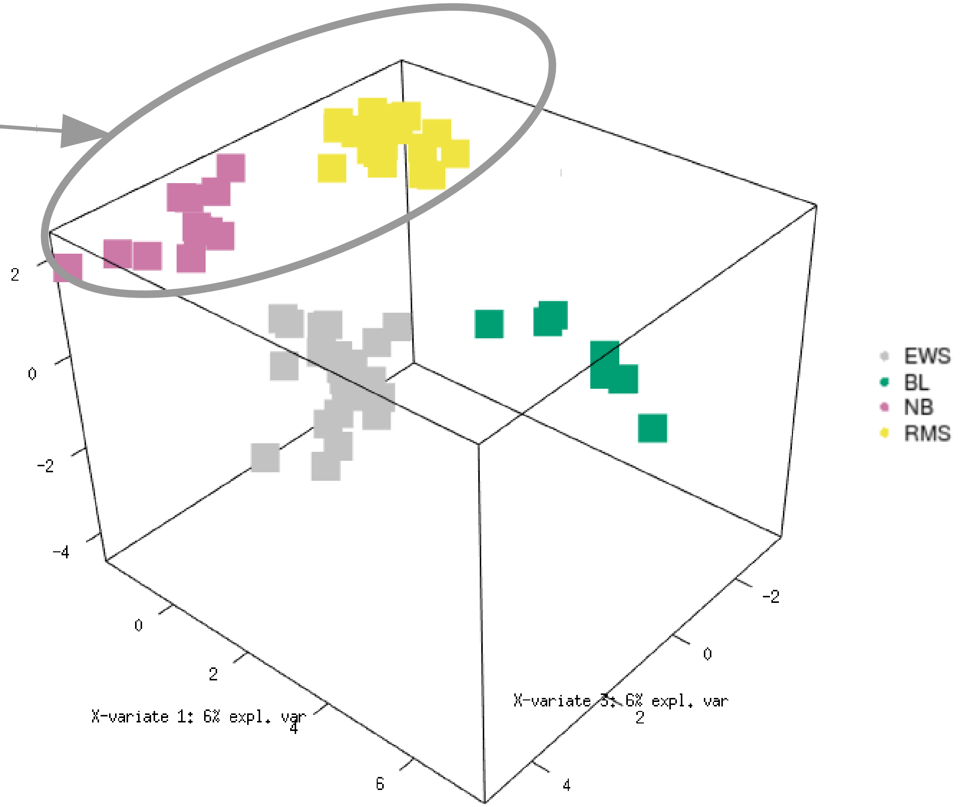


Individuals plot

Sparse PLS-DA

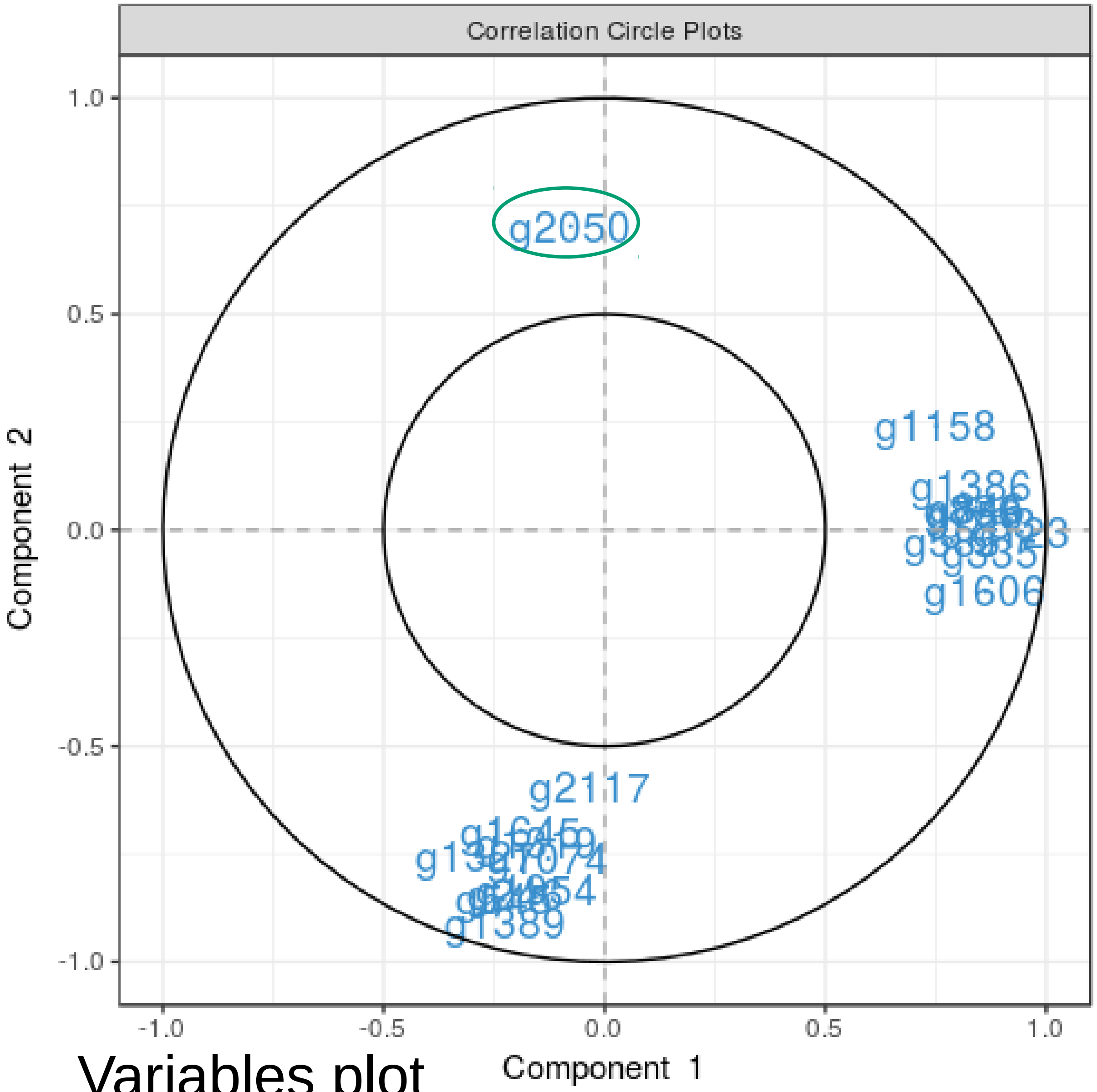



sPLS-DA on SRBCT, comp 1 & 2
X-variate 2: 5% expl. var
2.5
0.0
-2.5
0.0
2.5
5.0
7.5
X-variate 1: 6% expl. var
Legend
EWS
BL
NB
RMS

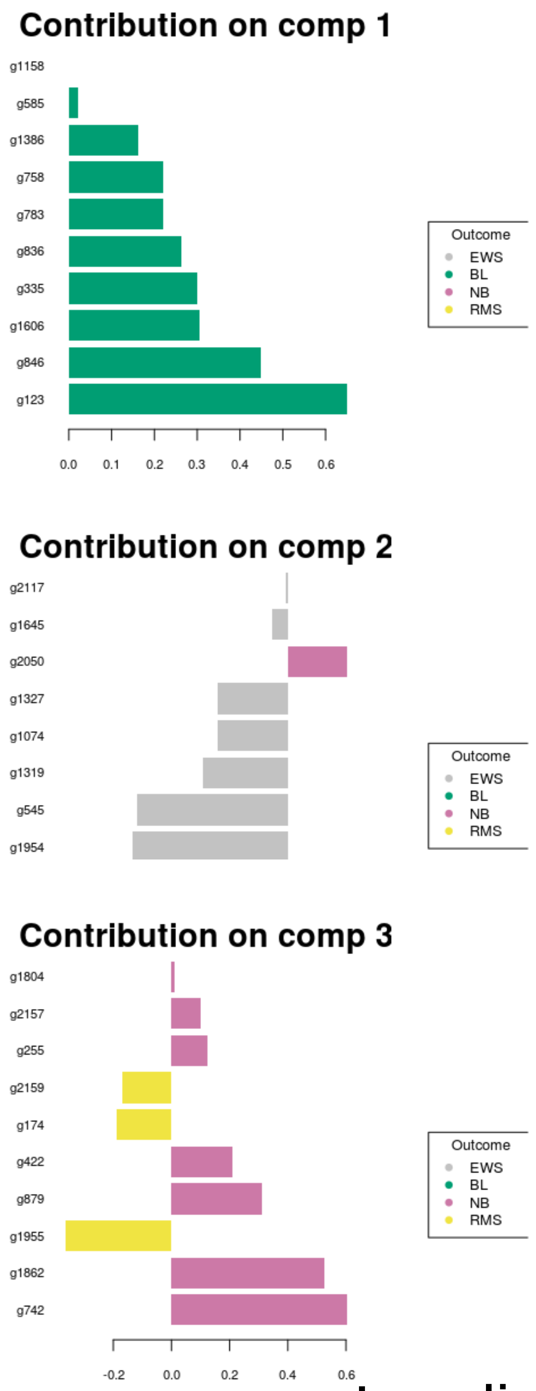


Individuals plot

Sparse PLS-DA

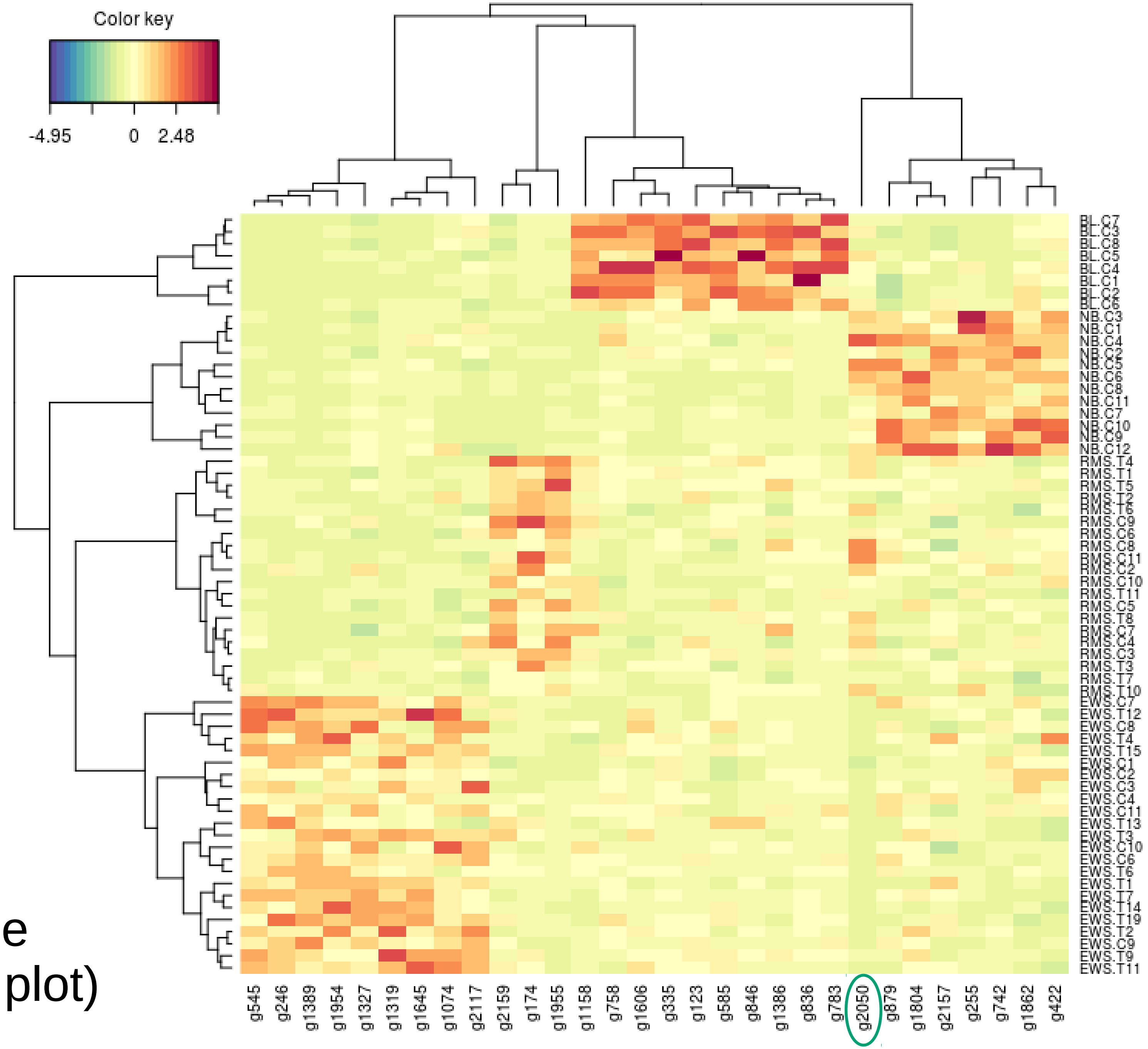


Variables plot



Loadings plot

Sparse PLS-DA



Clustered Image Map (variables plot)

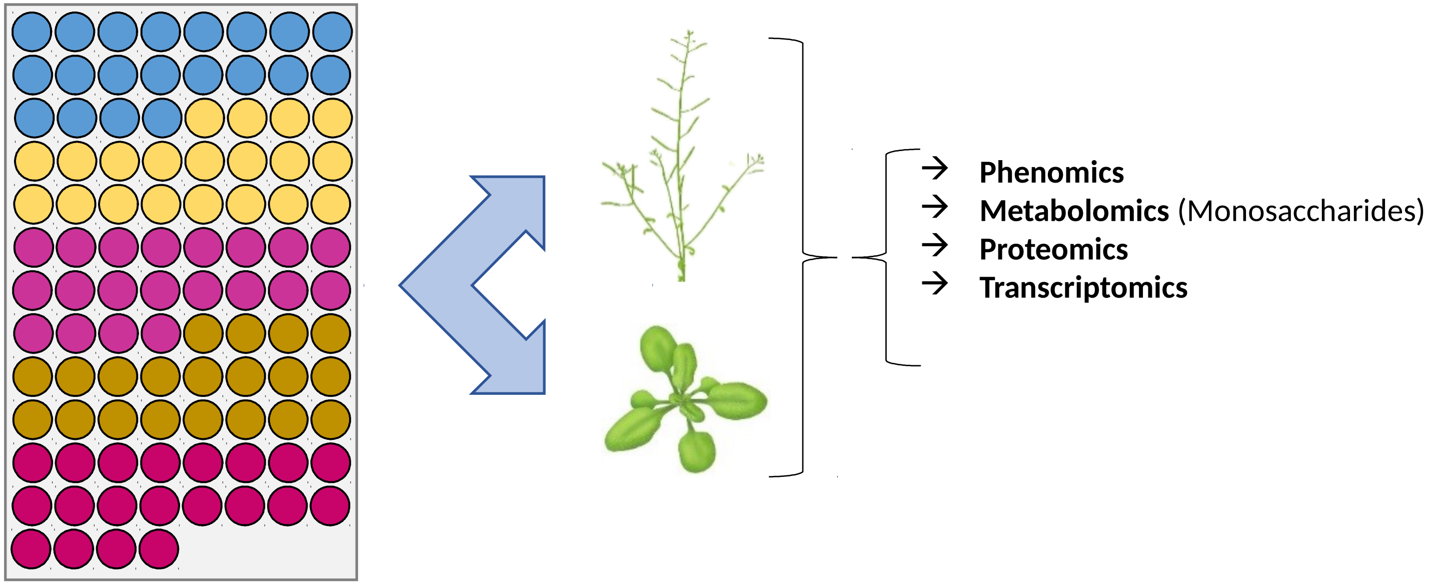
WallOmics

Experimental design

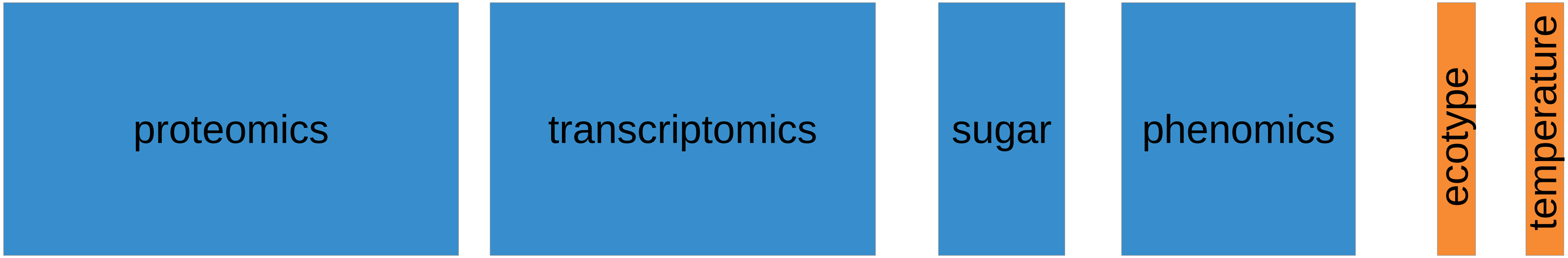
RSV Laboratoire de Recherche en Sciences Végétales

Harold Duruflé, Christophe Dunand, … Univ. Paul Sabatier LRSV

- **30** samples A. thaliana: **5** ecotypes x **2** temperatures x **3** replicates
- **4** data sets: phenomics, metabolomics (sugar), proteomics, transcriptomics



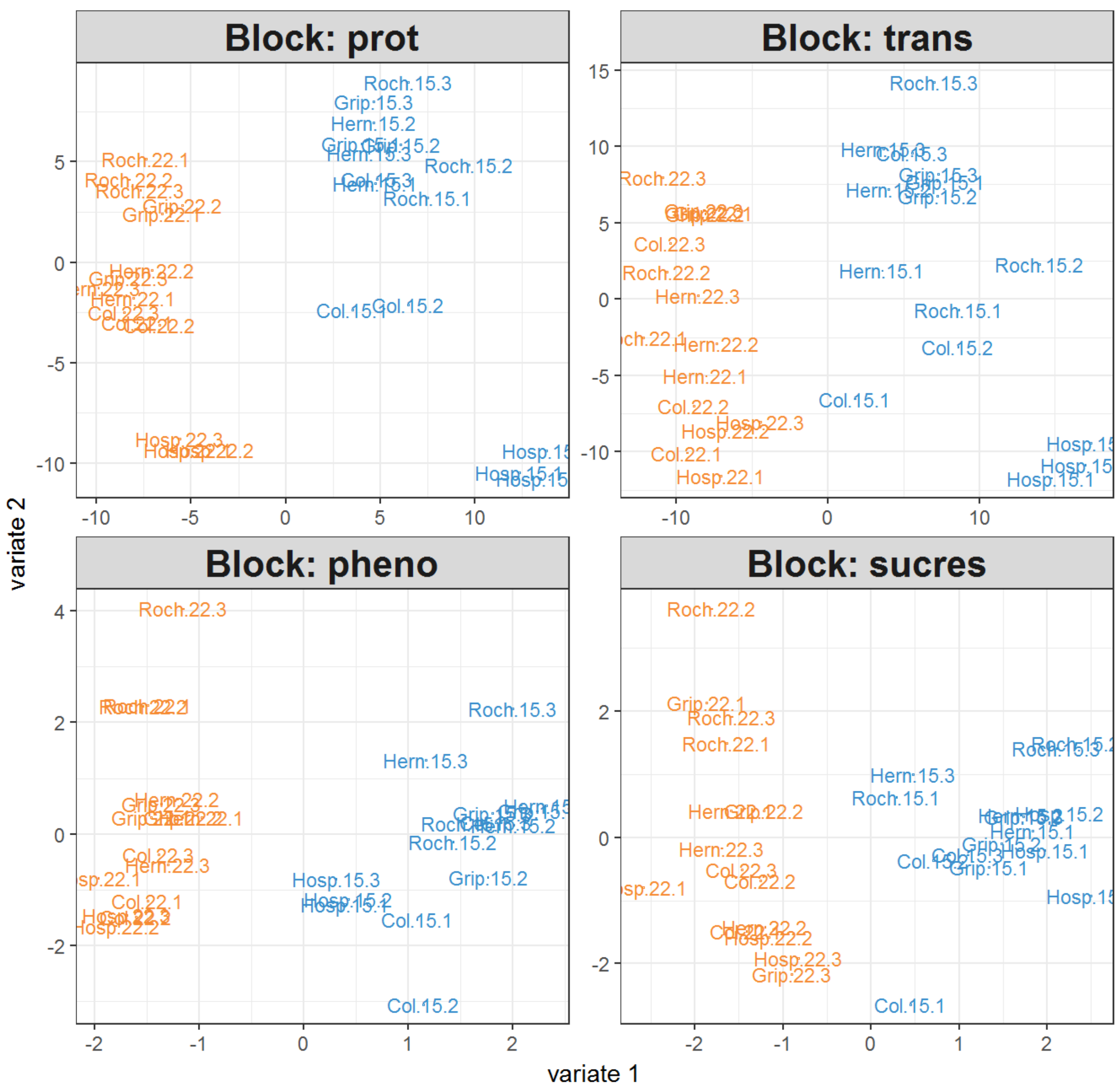
→ **Phenomics**
→ **Metabolomics** (Monosaccharides)
→ **Proteomics**
→ **Transcriptomics**



Supervised multi-block analysis

Factor: temperature

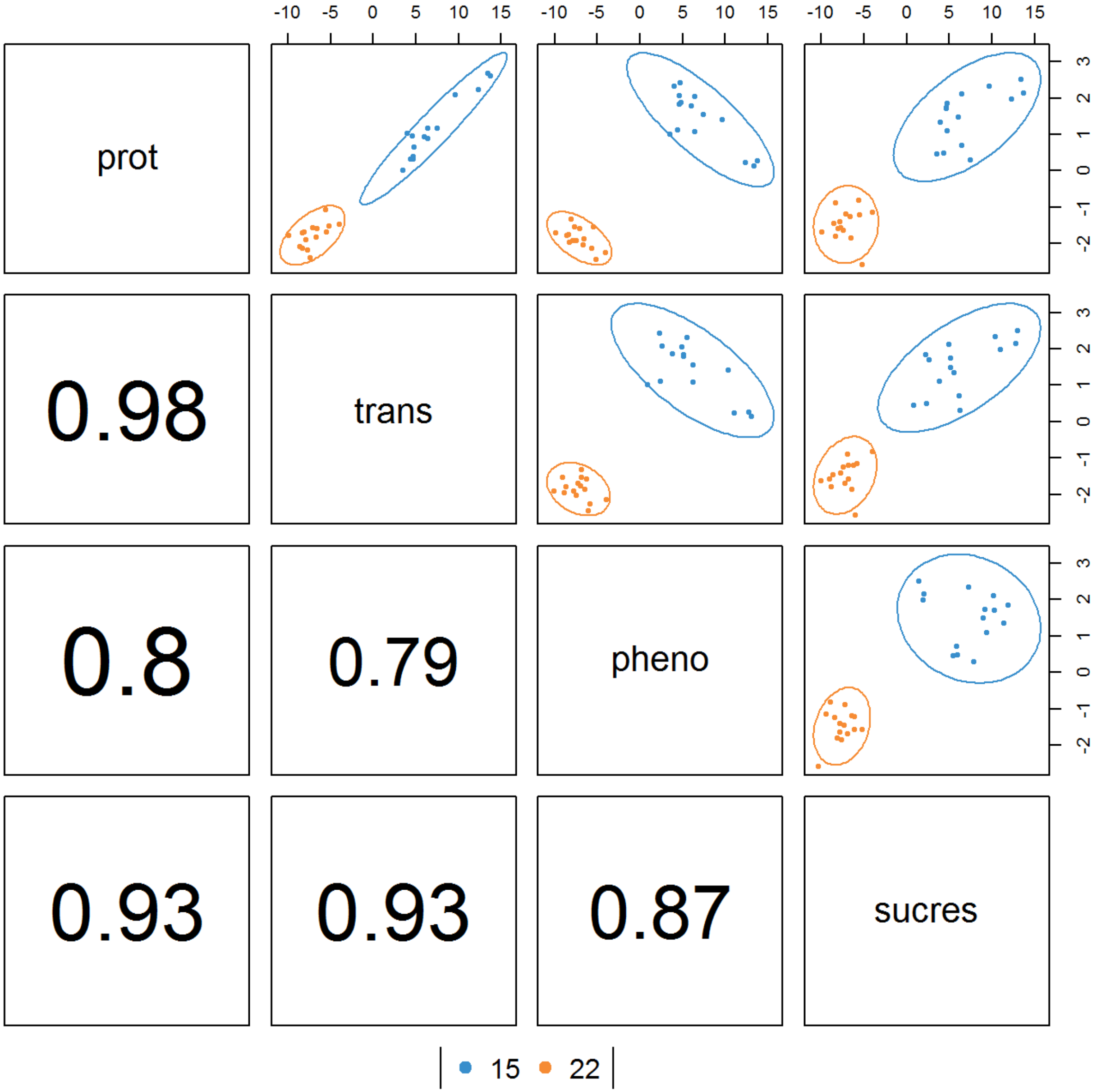
Individual plots



Supervised multi-block analysis

Factor: temperature

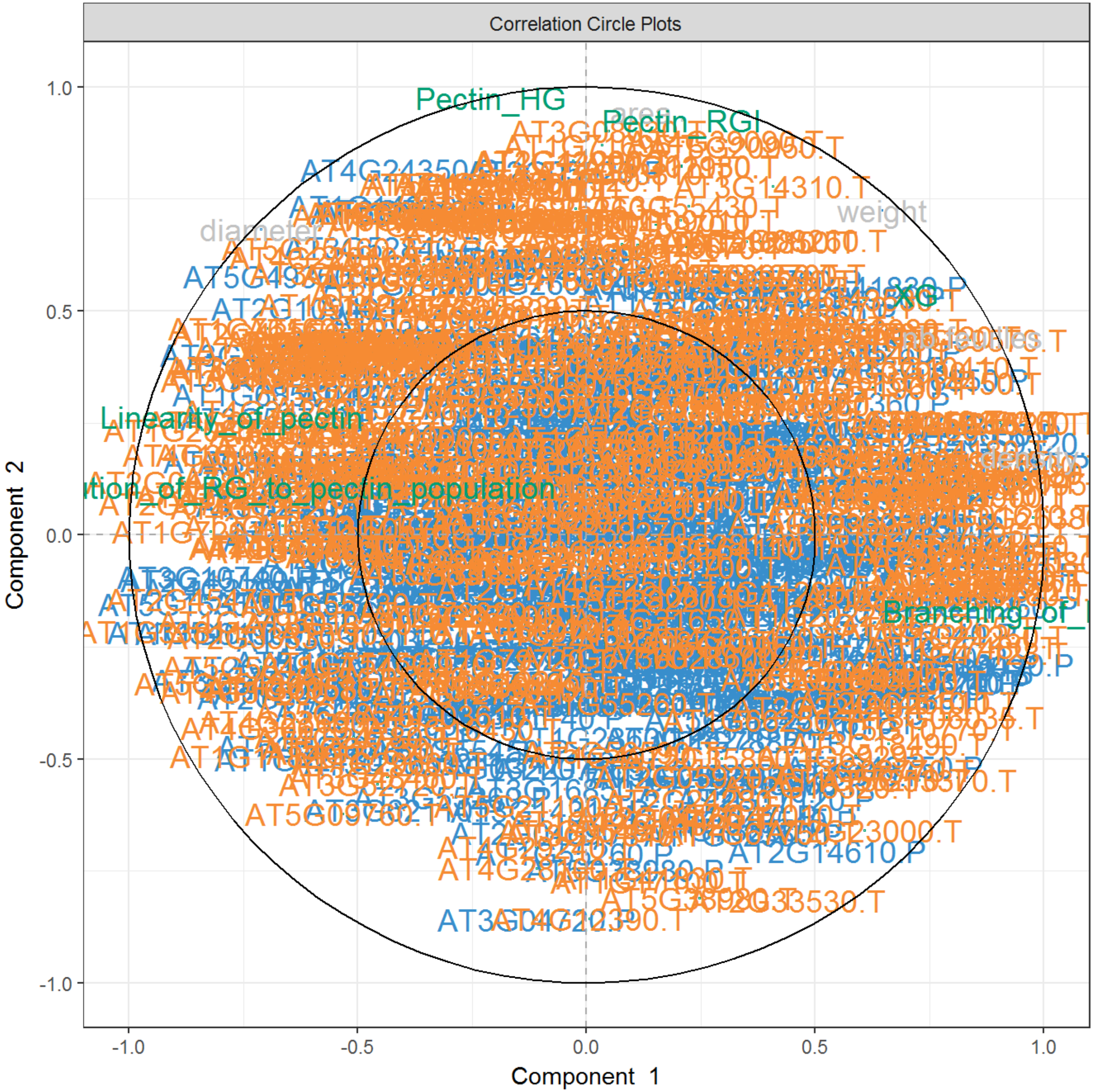
First components



Supervised multi-block analysis

Factor: temperature

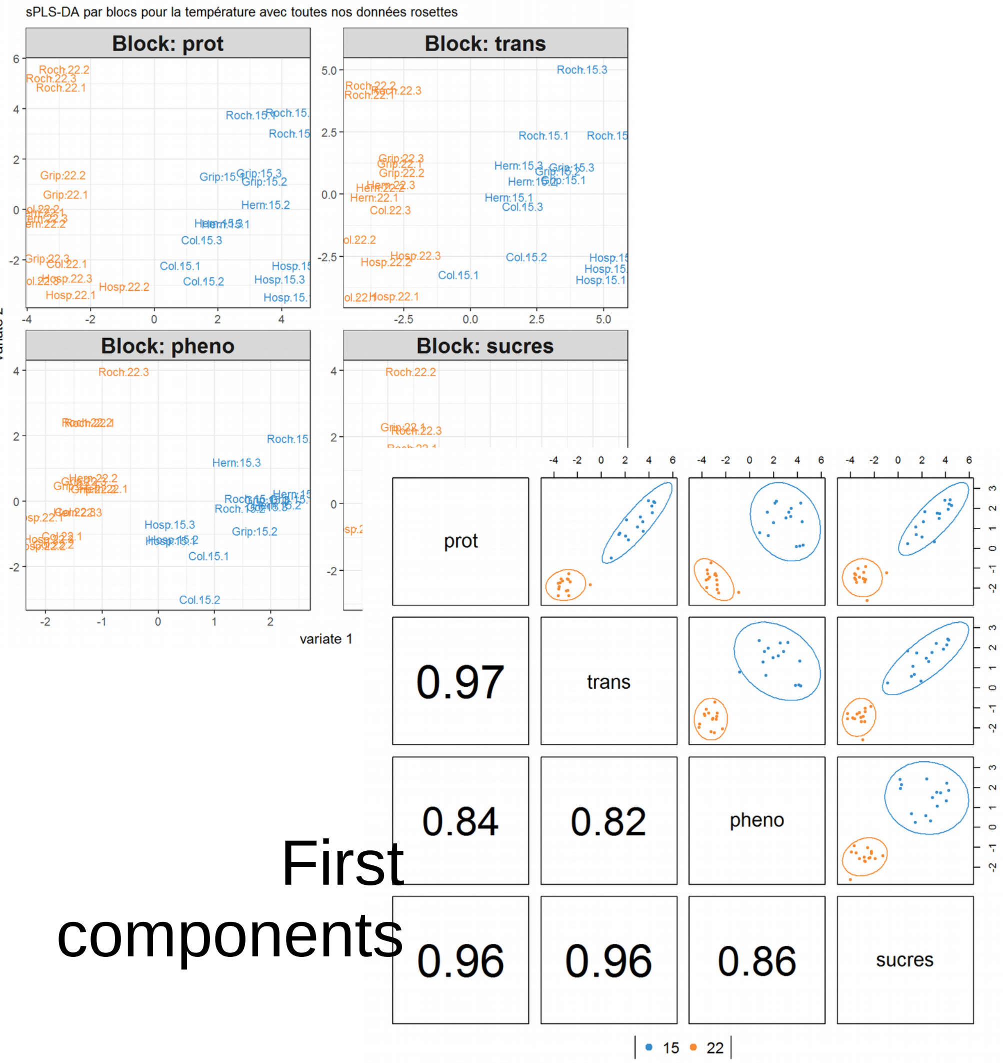
Variables plot



Supervised sparse multi-block analysis

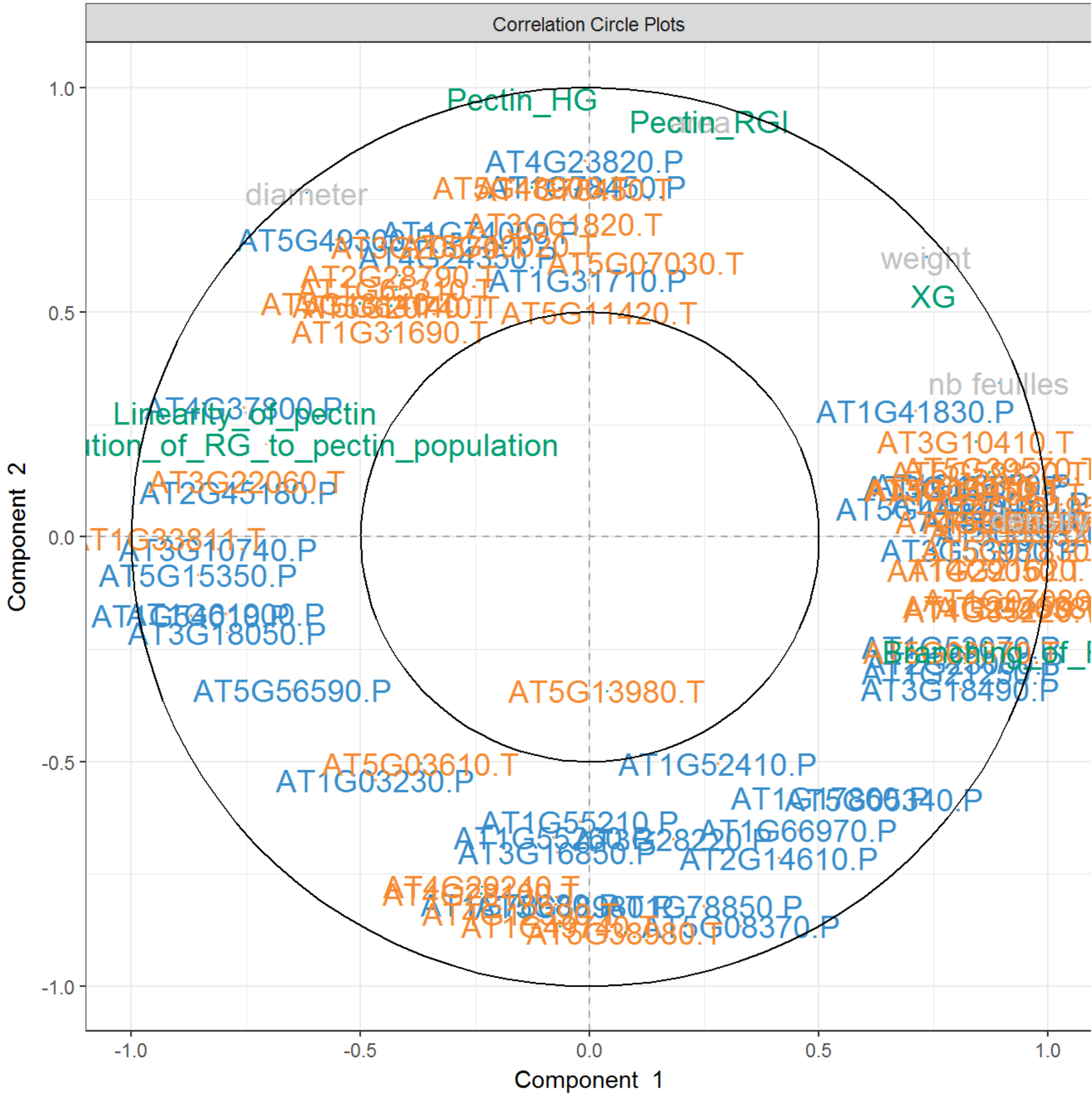
Factor: temperature

Individuals plot

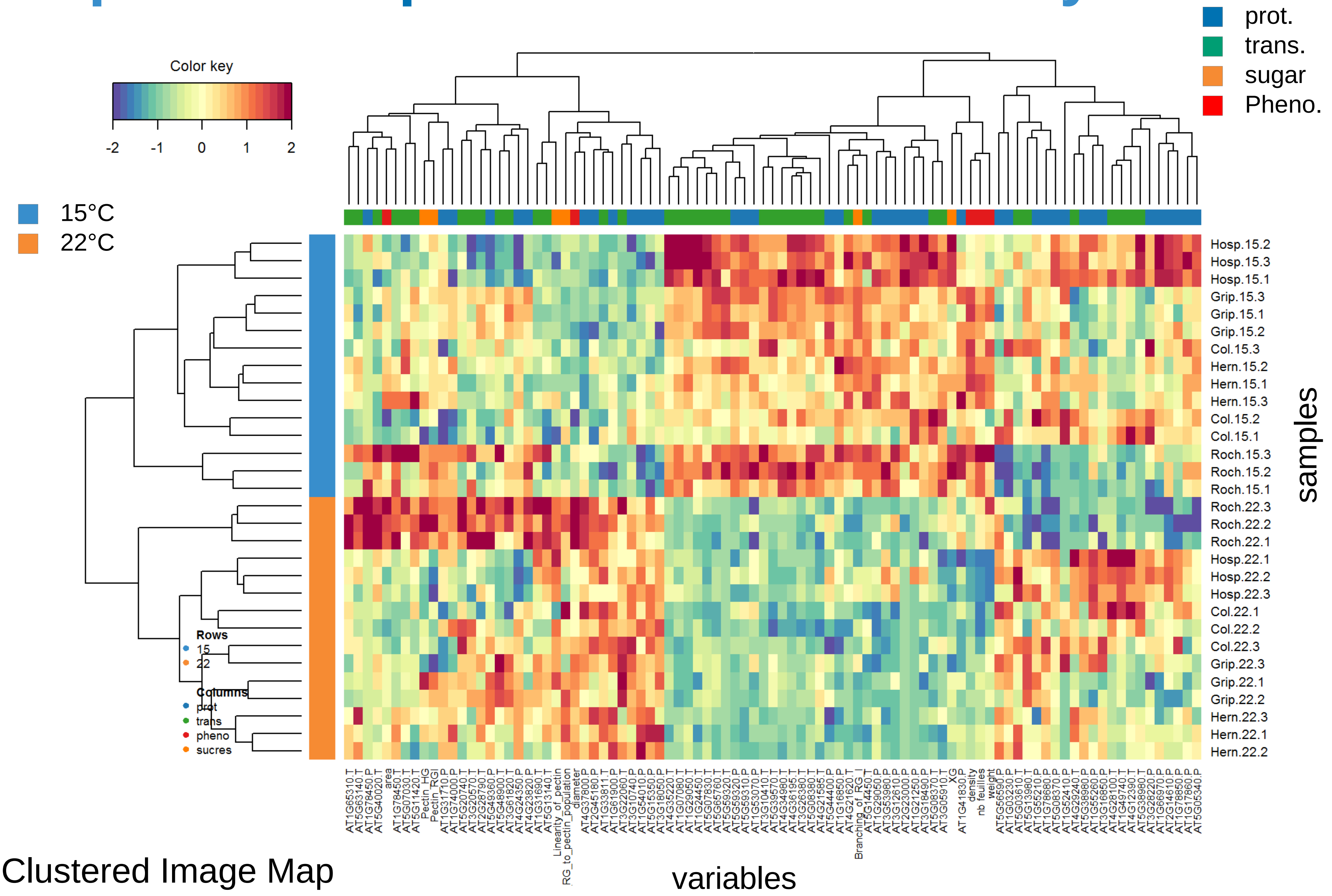


First components

Variables plot

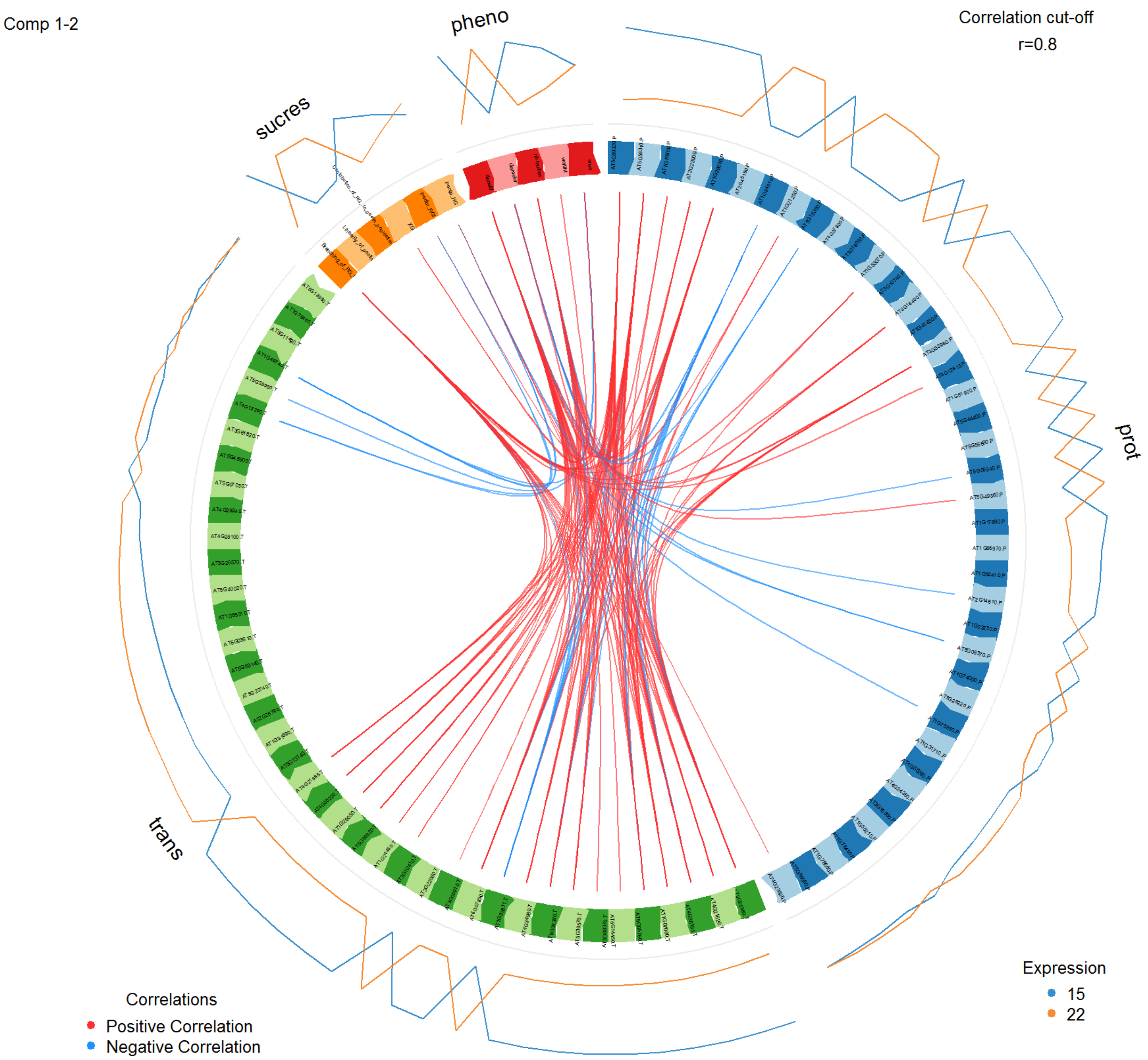


Supervised **sparse** multi-block analysis



Clustered Image Map

Supervised sparse multi-block analysis

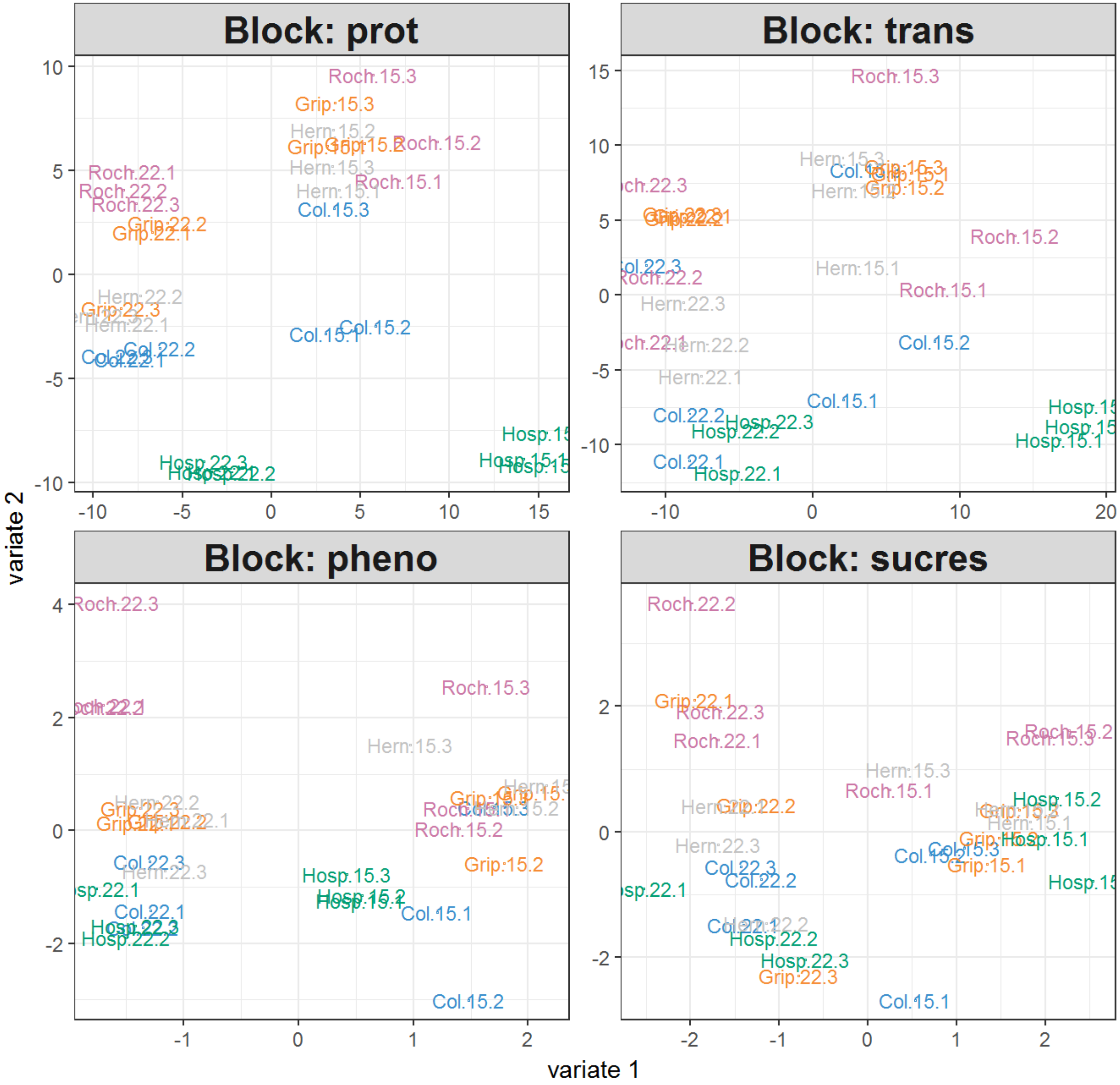


Circos plot

Supervised multi-block analysis

Factor: ecotype (5 categories)

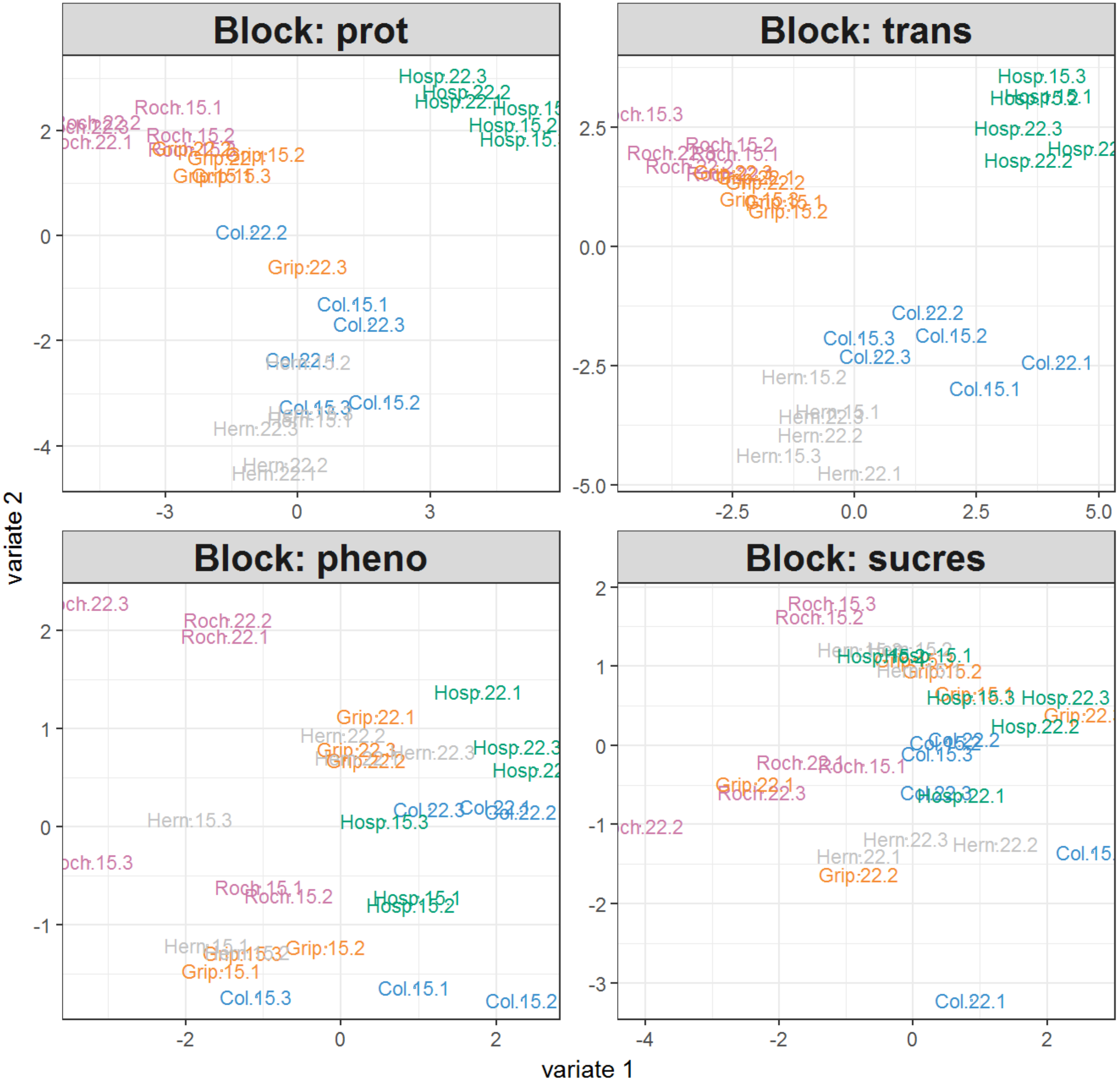
Individual plots



Supervised sparse multi-block analysis

Factor: ecotype (5 categories)

Individual plots

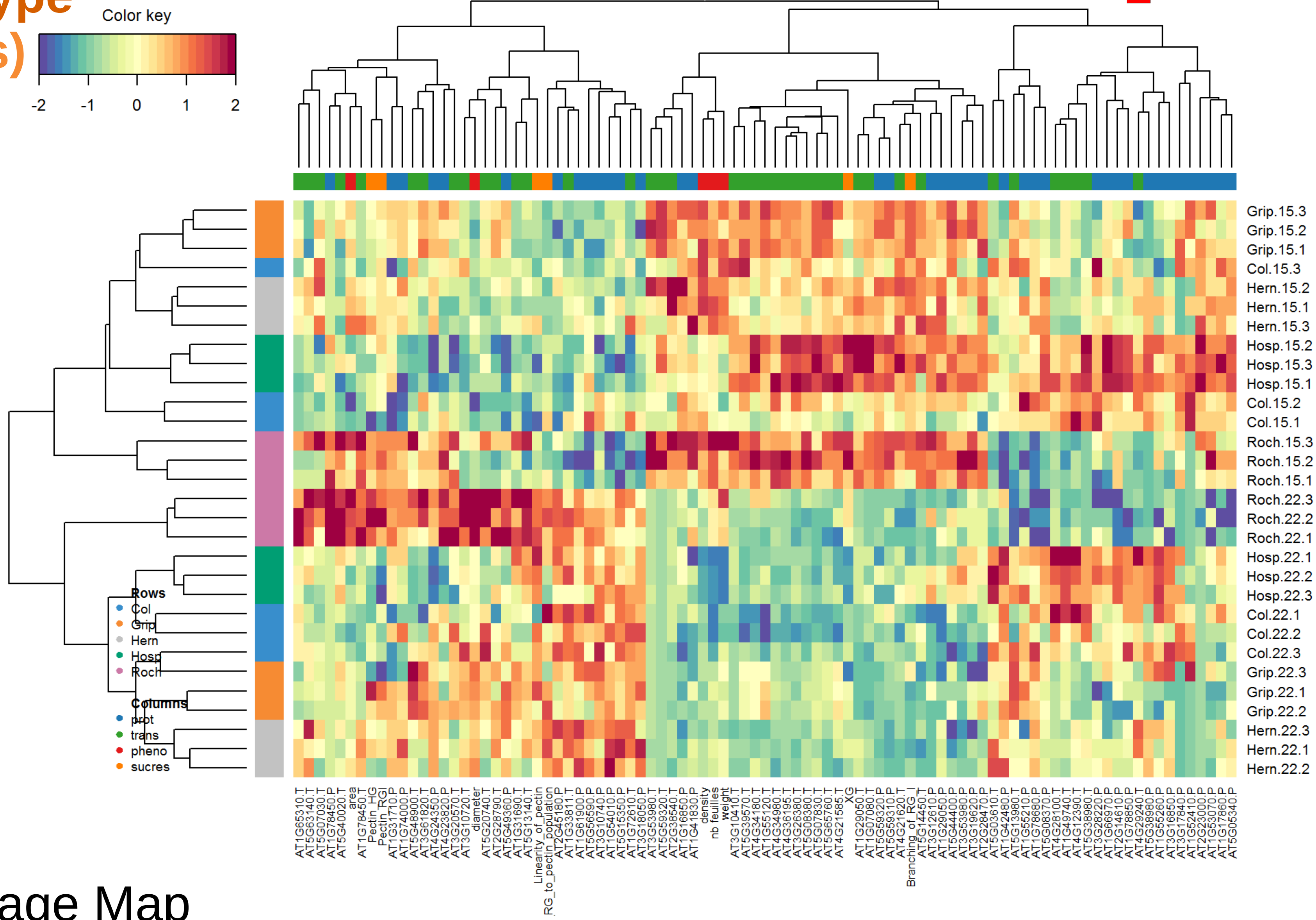


Supervised multi-block analysis

- 🟦 prot.
- 🟩 trans.
- 🟧 sugar
- 🟥 Pheno.

Factor: ecotype

(5 categories)

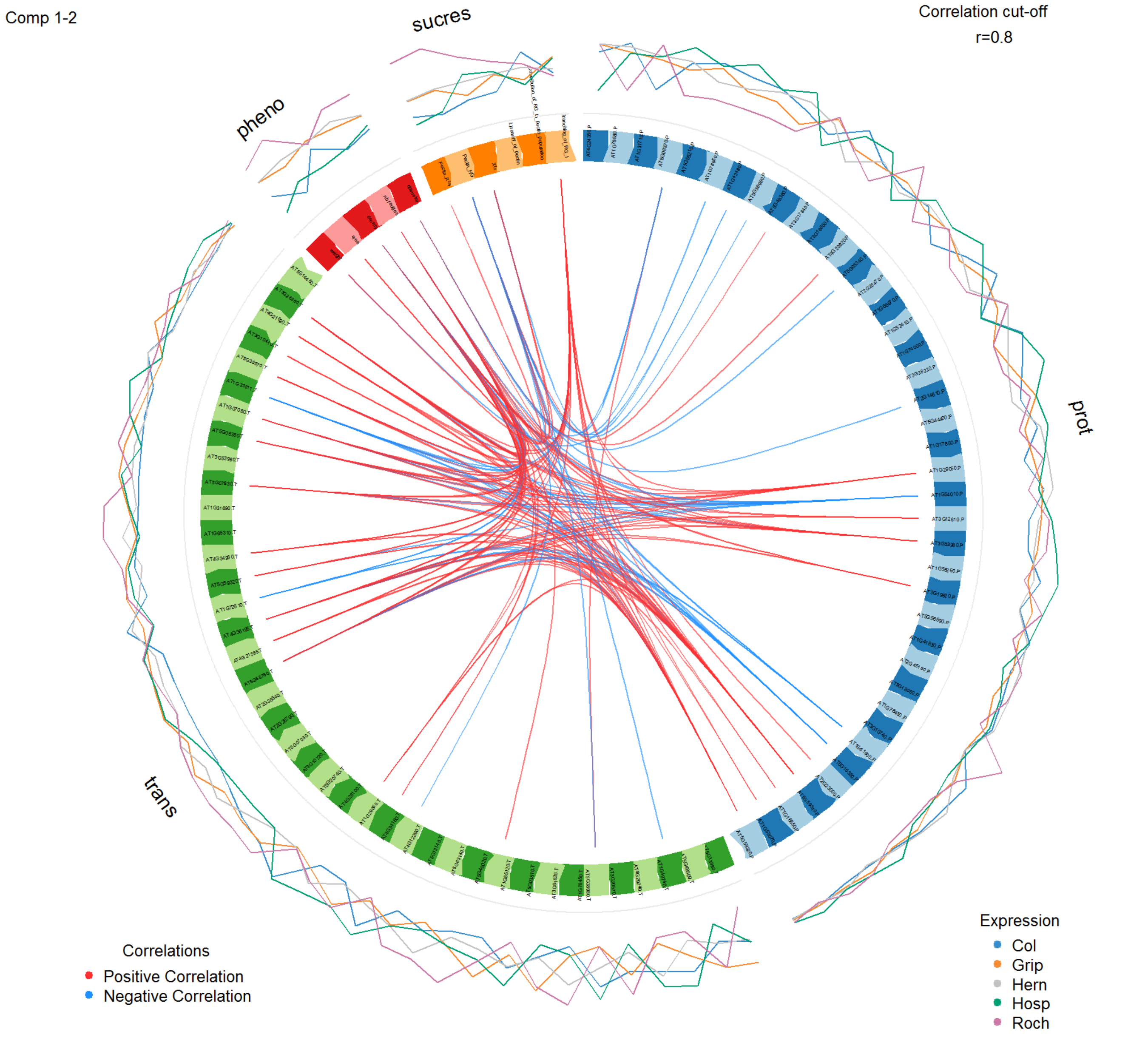


- 🟦 Col
- 🟩 Hosp
- 🟧 Grip
- ⬜ Hern
- 🟪 Roch

Clustered Image Map

Supervised multi-block analysis

Factor: ecotype
(5 categories)



Circos plot

Hints...

- 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 one data set)
- Address problems explicitly formulated: `` I *want to integrate my data*” is not a problem explicitly formulated
- Clearly identify supervised and unsupervised question and methods to use. ``PCA is not a good method, I can’t see my clusters...”