



HAL
open science

A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis

Harold Duruflé, Merwann Selmani, Philippe Ranocha, Elisabeth Jamet, Christophe Dunand, Sébastien Dejean

► To cite this version:

Harold Duruflé, Merwann Selmani, Philippe Ranocha, Elisabeth Jamet, Christophe Dunand, et al.. A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis. *Briefings in Bioinformatics*, 2021, 22 (3), pp.bbba166. <10.1093/bib/bbaa166>. <hal-02921927v3>

HAL Id: hal-02921927

<https://hal.science/hal-02921927v3>

Submitted on 3 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

1 **A powerful framework for an integrative study with heterogeneous omics data: from**
2 **univariate statistics to multi-block analysis**

3

4 Harold Duruflé^{1,2}, Merwann Selmani¹, Philippe Ranocha¹, Elisabeth Jamet¹, Christophe
5 Dunand^{1*}, Sébastien Déjean^{3*}

7 ¹ Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 24
8 chemin de Borde Rouge, Auzeville, BP 42617, 31326 Castanet-Tolosan, France.

9 ² LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France.

10 ³ Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, UPS, 31062
11 Toulouse, France

12 * Corresponding authors:

13 Sébastien Déjean (sebastien.dejean@math.univ-toulouse.fr; +33 (0)5 61 55 69 16)

14 Christophe Dunand (dunand@lrsv.ups-tlse.fr; +33 (0)5 34 32 38 58)

15

16

17 **ABSTRACT**

18 The high-throughput data generated by new biotechnologies used in biological studies require
19 specific and adapted statistical treatments. In this work, we propose a novel and powerful
20 framework to manage and analyse multi-omics heterogeneous data to carry out an integrative
21 analysis. We illustrate it using the package mixOmics for the R software as it specifically
22 addresses data integration issues. Our work also aims at confronting the most recent
23 functionalities of mixOmics to real data sets because, even if multi-block integrative
24 methodologies exist, they still have to be used to enlarge our know-how and to provide an
25 operational framework to biologists. Natural populations of the model plant *Arabidopsis*
26 *thaliana* are employed in this work but the framework proposed is not limited to this plant and
27 can be deployed whatever the organisms of interest and the biological question. Four omics
28 data sets (phenomics, metabolomics, cell wall proteomics and transcriptomics) have been
29 collected, analysed and integrated in order to study the cell wall plasticity of plants exposed to
30 sub-optimal temperature growth conditions. The methodologies presented start from basic
31 univariate statistics and lead to multi-block integration analysis, and we highlight the fact that
32 each method is associated to one biological issue. Using this powerful framework led us to
33 novel biological conclusions that could not have been reached using standard statistical
34 approaches.

35

36 Keywords: abiotic stress, *Arabidopsis thaliana*, integrative analysis, statistical framework,
37 systems biology.

38 1. INTRODUCTION

39 Biological processes can be studied using measurements that are ever more complex.
40 Today, biologists have access to plethora of new technologies to address their questions. The
41 high-throughput measurements have revolutionized the way to evaluate and predict the
42 behavior of organisms for example in response to environmental changes. Nowadays, one
43 biological sample can deliver many types of “big” data, such as genome sequences
44 (genomics), genes and proteins expression levels (transcriptomics and proteomics), metabolite
45 profiles (metabolomics) and phenotypic observations (phenomics). The revolution of high
46 throughput technologies has also greatly reduced the cost of those omics data production,
47 opening new prospects to the development of tools for data treatment and analysis (Li, Wu, &
48 Ngom, 2016; Meng et al., 2016).

49 The heterogeneous data collected from cellular to organism levels are associated to a
50 wide variety of techniques sometimes species-specific. The acquisition of data requires a
51 particular experimental design and a suitable methodology to highlight their mining (Rai,
52 Saito, & Yamazaki, 2017). An experimental design inadequate for an integrative analysis
53 could complicate the final interpretation of the collected data. On the contrary, a suitable
54 methodology of analysis can be optimized and brings keys to improve the visibility of the
55 whole data. This point of view was previously stated in (Kerr, 2003) for microarray studies:
56 *“While a good design does not guarantee a successful experiment, a suitably bad design*
57 *guarantees a failed experiment—no results or incorrect results”*.

58 Use of multi-omics data makes possible a deeper understanding of a biological system
59 (Zargar et al., 2016, Rajasundaram & Selbig, 2016). Indeed, quantification technologies
60 improve accuracy and create great potential for elucidating new questions in biology.
61 However, this technological revolution must be carefully used because the correlation
62 between quantification analyses is not effective. For example, it is known that it is usually

63 difficult to correlate transcriptomic and proteomic data (Duruflé et al., 2017; Jamet et al.,
64 2009; Maier, Güell, & Serrano, 2009). Each of these technologies has its own limitations and
65 collecting different types of data should help understanding the effects of one or more
66 experimental conditions. A cohort of hypotheses can be proposed with multi-omics analysis.
67 Thus, biological candidates can be identified as biomarkers (*e.g.* genes, proteins, molecules)
68 under complex environmental conditions, and/or new complex regulations can be found.

69 Altogether, it is generally admitted that studying a single kind of omics data is not
70 sufficient to understand the effects of a treatment on a complex biological system. To obtain a
71 holistic view, it is preferable to combine multiple omics analyses. To highlight the interest of
72 such integrative approaches, let us consider a toy example with two variables (V_x and V_y)
73 measured on 12 individuals (6 from one group called Controlled, and 6 from another one
74 called Treated). The values are presented in supplemental Table S1. Statistical tests
75 (Wilcoxon rank sum test and Student t test) do not reveal any significant difference between
76 the two groups for both the V_x and V_y variables when they are analysed separately (p-values
77 higher than 0.3). But, a simple scatterplot (Figure 1) highlights the interest of combining the
78 two variables. Indeed, it clearly appears that the two groups are separated if we consider the
79 V_x and V_y variables together.

80 Thus, in the same vein, we claim that the integrative analysis of several data sets
81 acquired on the same individuals can reveal information that single data set analysis would
82 keep hidden. Furthermore, the toy example also highlights the interest of a relevant graphical
83 representation: information hidden in supplemental Table S1 is clearly visible in Figure 1.
84 The recent work by (Matejka & Fitzmaurice, 2017) is assuredly a good way to be strongly
85 convinced about data visualisation.

86 This article focuses on a powerful framework we propose to manage and analyse
87 heterogeneous data sets acquired on the same samples. It proceeds step by step, from basic

88 univariate statistics to multi-block integration analysis (Singh et al., 2016; Tenenhaus et al.,
89 2014). We illustrate the gaps bridged by each method from the computation of univariate
90 statistics to a thorough implementation of multi-block exploratory analysis. The
91 implementation of the methods and the graphical visualizations have simply been
92 accomplished with existing tutorials for the R software (R Core Team, 2018) and the
93 mixOmics package (Rohart et al., 2017). But, since their interpretation is not easy (González
94 et al., 2013), this article will provide a better understanding of the statistical integration and a
95 way to include it in a global reflexion structured in a workflow summarized in Figure 2. We
96 also aim at increasing our know-how related to these novel methodologies by confronting
97 them to new real data sets. The first section presents the background of our study and the data
98 sets we have dealt with detailed in (Duruflé, 2019a; 2019b). Then, we describe several
99 statistical methods used to address specific biological questions. Afterwards, we explain in
100 detail the statistical results and give clues to interpret them.

101

102 **2. BIOLOGICAL CONTEXT**

103 In the global warming context, seasons are altered with modifications of the
104 temperatures. The elevation of the temperature is the most studied change because it is
105 already observed (Savo et al., 2016). The occurrence of cold stress can also appear without
106 any previous chilling period and it could become a problem to maintain agricultural
107 productivity in the future (Gray & Brady, 2016). The model plant *Arabidopsis thaliana* of the
108 *Brassicaceae* family has a worldwide geographical distribution and therefore has to adapt to
109 multiple and contrasted environmental conditions (Hoffmann, 2002). The huge accumulation
110 of molecular data concerning this plant is very helpful for studying complex multiple levels
111 responses. It is expected to transfer obtained results to other plant species of economic interest
112 for translational pipelines (Sibout, 2017).

113

114 2.1. Experimental design

115 First, a compromise is necessary to determine the ideal number of biological
116 replicates. It is hard to find an agreement between the reality of the biological experimentation
117 (*e.g.* limitation in material, space, time, work force and cost) and the necessity to get robust
118 information for the statistical analyses. The method used for the randomization of the
119 replicate also needs to be considered. For these reasons, the experimental protocol must
120 minimize potential external impacts within and between the replicates and avoid confounded
121 effects.

122 To strengthen the results, each biological replicate can be the average of several
123 technical replicates, if the type of analysis allows it. For the biologist, it is important to know
124 the number of experimental repetitions to appreciate the variability between the different
125 conditions. But, for a statistician, the information resides into the intrinsic variability of the
126 different samples or repetitions. For all these reasons, one sample considered as “out of
127 norms” by the biologist could be valuable in a multi-omics analysis.

128 Our experimental design was built with two crossed factors: i) ecotypes with 5 levels
129 (4 Pyrenees Mountain ecotypes Roch, Grip, Hern, Hosp, living at different altitudes, and Col,
130 a reference ecotype from Poland, living at low altitude) and ii) temperature with 2 levels
131 (22°C and 15°C). For each ecotype, rosettes and floral stems were collected and analysed. At
132 22°C, rosettes were collected at 4 weeks, *i.e.* at the time of floral stem emergence. At 22°C,
133 floral stems were collected at 6, 7 and 8 weeks respectively for Col, Roch / Grip and Hern /
134 Hosp. At 15°C, rosettes and stems were collected 2 weeks later than at 22°C. More details
135 about the plant culture conditions can be found in (Duruflé, 2019a). Three independent
136 biological replicates were analysed for each sample including 20 plants per sample. To

137 minimize the experimental effect, each plant was grown at a randomly chosen place according
138 to the experimental design represented in Figure 3.

139

140 2.2. Omics data sets and curation

141 In this project, the four following omics data sets (called blocks thereafter) were collected:

142 (i) Phenomics, *i.e.* a macro phenotyping analysis, was performed on two organs: rosettes
143 and floral stems (Duruflé, 2019a). Indeed at the time of sample collection and prior to
144 freezing, 9 phenotypic variables were measured: 5 on rosettes (mass, diameter,
145 number of leaves, density, and projected rosette area), and 4 on floral stems (mass,
146 diameter, number of cauline leaves, length).

147 (ii) Metabolomics, *i.e.* identification and quantification of seven cell wall
148 monosaccharides (fucose, rhamnose, arabinose, galactose, glucose, xylose and
149 galacturonic acid), were performed as previously described (Duruflé et al., 2017).
150 Theoretical cell wall polysaccharide composition was inferred, based on the
151 monosaccharide analyses according to (Duruflé et al., 2017; Houben et al. 2011;
152 Duruflé, 2019a).

153 (iii) Proteomics, *i.e.* identification and quantification of cell wall proteins by LC-MS/MS
154 analyses, were performed as described (Duruflé et al., 2017). Altogether, 364 and 414
155 cell wall proteins (CWPs) were identified and quantified in rosettes and floral stems,
156 respectively (Duruflé, 2019b).

157 (iv) Transcriptomics, *i.e.* sequencing of transcripts also called RNA-seq, was performed
158 according to the standard Illumina protocols as described (Duruflé et al., 2017).
159 Altogether, 19763 and 22570 transcripts were analysed in rosettes and floral stems,
160 respectively (Duruflé, 2019b).

161

162 3. TIDYING DATA

163 Statistical data analysis requires efficient data pre-processing. As mentioned in
164 (Wickham, 2014), “*It is often said that 80% of data analysis is spent on the process of*
165 *cleaning and preparing the data*”. So in an integrative analysis framework, each data set
166 needs to be structured in the same way, and (Wickham, 2014) has also stressed the following
167 statements: *1/ Each variable forms a column. 2/ Each observation forms a row*. So, in our
168 context, each data set is structured with biological samples in rows and variables in columns.

169 Handling missing data is always a big deal. As stated by Gertrude Mary Cox (an
170 American statistician of the 20th century), “*the best thing to do with missing values is not to*
171 *have any*”. Fortunately, many methods exist to deal with missing values. For instance, the
172 methodologies implemented in the missMDA package (Husson & Josse, 2013) are dedicated
173 to the handling of missing values in the context of multivariate data analysis. For example in
174 this work, missing proteomics quantification data were dealt with considering two situations:
175 (i) non-validated proteins (identification with a single specific peptide and/or in a single
176 biological replicate); and (ii) undetectable proteins (no peptide identified in a given
177 condition). In the former case, a background noise, corresponding to the minimum, and the
178 first statistical quartile of the biological replicate, was applied. In the latter case, a background
179 noise of 6 (value lower than the minimum value found in the whole experiment) was applied.
180 This treatment allowed combining the quantification process with the qualitative study and
181 provided a higher confidence in the final result.

182 More recently, a study focused on missing rows in data sets in an integrative
183 framework (Voillet et al., 2016). Within an integrative study, we can easily be in this case if,
184 for instance, the number of biological replicates is not the same for transcriptomics and
185 proteomics analyses. The main idea to remember would be to deal with missing values with

186 an *ad-hoc* method taking into account the specificity of the data. In our case, two replicates of
187 the transcriptomic data had to be deleted due to their low quality. Following the method
188 proposed in (Voillet et al., 2016), these missing rows were imputed using the samples for
189 which all the data were available, *i.e.* the two other replicates.

190

191 **4. RATIONALE SUPPORTING THE PROPOSED FRAMEWORK**

192 4.1. Software

193 As mentioned in the Comprehensive R Archive Network (CRAN, cran.r-project.org),
194 *R “is a freely available language and environment for statistical computing and graphics*
195 *which provides a wide variety of statistical and graphical techniques: linear and nonlinear*
196 *modelling, statistical tests, time series analysis, classification, clustering, etc.”*

197 R functions with a command-line interface that, even if it can appear not user-friendly,
198 allows the user to build scripts that can be run on various data sets with rather few tuning. R
199 gives access to the newest methodological developments due to its very active community (R-
200 bloggers, R-help, UseR conference...) motivated by open science considerations. Furthermore,
201 efficient tools such as RStudio (www.rstudio.org) were developed in order to make the
202 initiation to R easier. In addition, many resources are available on CRAN to start with R.
203 Therefore, it seems highly reasonable to expect that the user can read, use and adapt existing
204 scripts available in the examples of each manual of packages after few hours of practice.
205 Specifically considering the community of biologists using R, the Bioconductor repository
206 (<http://www.rstudio.org/>) provides selected tools for the analysis of high-throughput genomic
207 data (Gentleman et al., 2004).

208 The dynamism around R appears in the packages developed by and for the
209 community. So, several packages exist to address statistical integrative studies. We focus on

210 the mixOmics package (Lê Cao et al., 2009; Rohart et al., 2017), but other packages such as
211 FactoMineR (Husson & Josse, 2013) can also be used for nearly similar purposes.
212 Methodologies presented in (Bécue-Bertaut & Pagès, 2008) and (Sabatier et al., 2013) are
213 also alternatives, as well as the Multi-Omics Factor Analysis (MOFA) approach proposed in
214 (Argelaguet et al., 2018). Regarding commercial software for instance, SIMCA-P (Umetrics,
215 umetrics.com/) propose several methods to perform integrative analyses, and toolboxes for
216 Matlab are also available (The MathWorks, Inc., Natick, Massachusetts, United States). We
217 choose to favor an open source software, as it is easier to promote a free software than a
218 commercial one when people are not specialists in the domain (Carey & Papin, 2018).
219 Furthermore, mixOmics appears as a very active package addressing data integration issues. It
220 has been downloaded more than 25,000 times (unique IP address) in 2017, 5 versions were
221 released in 2017, the reference article (Lê Cao et al., 2009) has been cited 300 times and the
222 mixOmics team has published 16 articles related to this package since 2008.

223

224 4.2. One purpose, one method

225 In this section, partly inspired from the tutorial of mixOmics (mixomics.org), we wish to
226 highlight the link between a biological question (purpose) and the appropriate statistical
227 method.

228 • *Purpose: explore one single quantitative variable (e.g. what is the level of expression*
229 *of one gene?). Method: univariate elementary statistics such as mean, median for main*
230 *trends, and standard deviation or variance for dispersion, can be completed with a*
231 *graphical representation such as boxplot.*

232 • *Purpose: assess the influence of one single categorical variable on a quantitative*
233 *variable (e.g. Are the plant growth different in two or more environmental*

234 *conditions?*). Method: statistical significance test such as Student t test or Wilcoxon
235 rank sum test for two groups and ANOVA or Kruskal-Wallis for more groups will
236 address this question (McDonald, 2009). In this context, a special attention must be
237 paid to the structure of the data: independent samples (*e.g.* independent groups
238 observed in various conditions) or paired samples (*e.g.* same samples observed twice
239 or more in various conditions).

240 • *Purpose: evaluate the relationships between two quantitative variables (e.g. Is there a*
241 *correlation between the concentration of one protein and its transcript abundance?).*
242 Method: correlation coefficients (Pearson for linear relationships and Spearman for
243 monotonous ones) (McDonald, 2009). Graphical representations of correlation
244 matrices can provide a global overview of pairwise indicators (Friendly, 2002;
245 Murdoch & Chow, 1996).

246 • *Purpose: explore a single data set (e.g. transcriptomics) and identify the trends or*
247 *patterns in the data, experimental bias or, identify if the samples ‘naturally’ cluster*
248 *according to the biological conditions (e.g. Can we observe the effect of different*
249 *environmental growth conditions on different ecotypes?).* Method: an unsupervised
250 factorial analysis such as Principal Component Analysis (PCA) (Mardia, Kent, &
251 Bibby, 1980) provides such information about one data set without any *a priori* on the
252 result. Centering and scaling the data, such that all variables have zero mean and unit
253 variance, before performing PCA is usually useful when dealing with omics data to
254 make the PCA results meaningful.

255

256 The previously mentioned methods are rather standard and usually used for biological
257 data analysis whereas the methods mentioned hereafter are less usual.

258 • *Purpose: classifying samples into known classes based on a single data set (e.g. Can*
259 *we classify various ecotypes according to their transcriptomics profile?). Method:*
260 *supervised classification methods such as Partial Least Square Discriminant Analysis*
261 *(PLS-DA) (Lê Cao, Boitard, & Besse, 2011) assess how informative the data are to*
262 *rightly classify samples, as well as to predict the class of new samples.*

263 • *Purpose: unravel the information contained in two data sets, where two types of*
264 *variables are measured on the same samples (e.g. What are the main relationships*
265 *between the proteomics and transcriptomics datasets?). Method: using PLS-related*
266 *methods (Wold et al. , 2001) enable knowing if common information can be extracted*
267 *from the two data sets (or highlight the relations between the two data sets).*

268

269 The following methods are very recent and few applications have been published so
270 far. This work contributes to improve their efficiency on real data sets.

271 • *Purpose: the same as above but considering more than two data sets (e.g. What are*
272 *the main relationships between the proteomics, transcriptomics and phenotypic*
273 *data?). Method: multi-block PLS related methods were recently developed to address*
274 *this issue (Günther et al., 2014; Singh et al., 2019).*

275 • *Purpose: the same as above but in a supervised context (e.g. Can we determine a*
276 *multi-omics signature to classify ecotypes?). Method: multi-block PLS-DA (referred*
277 *as DIABLO for Data Integration Analysis for Biomarker discovery using Latent*
278 *variable approaches for Omics studies) was recently developed to address this issue*
279 *(Singh et al., 2019).*

280

281 A schematic view of the data sets and the methods implemented is presented in Figure 4.
282 The way to perform an integrative statistical study is illustrated through several cycles (Figure
283 4B). We prefer this view rather than a straightforward pipeline beginning with univariate
284 analysis and ending with multi-block approaches. Each method contributes to the global
285 comprehension of the data and can challenge the others. For instance, univariate statistics may
286 highlight outliers or essential variables. On the other hand, multi-block approaches may focus
287 on new samples and/or variables showing specific behavior that should be studied through a
288 univariate method. We claim that, facing integrative studies, a relevant statistical analysis
289 must go through these cycles, with progress and feedback.

290

291 4.3. Sparse extensions

292 Every methods developed in mixOmics are proposed with a sparse extension (sparse
293 PCA (S-PCA), sparse PLS (S-PLS)...). Sparse methods are useful to remove non-informative
294 variables (*e.g.* which can be considered as background noise) regarding the purpose of the
295 multivariate method. Concerning PCA for instance, the sparse version selects only the
296 variables that highly contribute to the definition of each principal component (PC), removing
297 the others. Sparsity is mathematically achieved via Least Absolute Shrinkage and Selection
298 Operator (LASSO) penalizations (Tibshirani, 1996).

299 In practice, the use of sparse methods in the context of omics data is very useful as it
300 reduces the number of potentially relevant variables displayed on the graphical outputs. Thus,
301 it facilitates the biological interpretation of the results and minimizes the list of potential
302 candidates for further investigations.

303

304 4.4. Numerical and graphical outputs

305 As previously mentioned, statistical analysis should be associated with graphical
306 representations (Figure 4C). A famous sentence assigned to Francis John Anscombe (a British
307 statistician of the 20th century) emphasized this point of view: “... *make both calculations and*
308 *graphs. Both sorts of output should be studied; each will contribute to understanding.*”
309 (Anscombe, 1973). Based on this principle, a recent work by Matejka and Fitzmaurice
310 (Matejka & Fitzmaurice, 2017) illustrates in a quite funny way how same numerical outputs
311 can provide very different graphical representations (including a scatterplot looking like a
312 dinosaur named *datasaurus*).

313 The results of univariate and bivariate approaches are mainly reported as p-values for
314 statistical testing. Boxplots and barplots, as produced, for instance, by the ggplot2 package
315 (Wickham, 2016), may complete and reinforce the interpretation of the results (Figure 4C).
316 Regarding barplots, one core question relies on the error bars that are frequently added:
317 should they be based on standard deviation or on standard error of the mean? A thorough
318 explanation about the difference is provided in (Cumming, Fidler, & Vaux, 2007). The
319 authors also mention this statement that may seem obvious but that is sometimes forgotten:
320 “*However, if n is very small (for example n = 3), rather than showing error bars and*
321 *statistics, it is better to simply plot the individual data points.*”

322 We also used graphical representations of correlation matrices (Figure 4C) such as
323 those produced by the corrplot package (Wei & Simko, 2016) for the R software. This is
324 essential when dealing with (not so) many variables: with 50 variables, 1225 (50 x 49 / 2)
325 pairwise correlation coefficients are computed and have to be analysed and interpreted.

326 Regarding multivariate analyses (from PCA to multi-block analyses), we used the
327 graphical outputs provided by the mixOmics R package (Rohart et al., 2017). They are based
328 on the representation of individuals and variables projected on specific sub-spaces (Figure

329 4C). A thorough discussion about the complementarity between several graphical displays is
330 given in (González et al., 2013).

331 In a multivariate supervised analysis, the individuals (biological samples) of the study
332 are represented as points located in a specific sub-space defined by the first PLS-components
333 (Figure 4C). Interpretation is based on the relative proximities of the samples and on the
334 equivalent representation for variables.

335 The standard representation for the variable plots is frequently referred as correlation
336 circle plot (Figure 4C). It was primarily used for PCA to visualise relationships between
337 variables, but it has been extended to deal with multi-block analysis. In such a plot, the
338 correlation between two variables can be visualised through the cosine of the angle between
339 two vectors starting at the origin and ending at the location of the point representing the
340 variable. The representation of variables can also be done through a relevance network. These
341 networks are inferred using a pairwise similarity matrix directly obtained from the outputs of
342 the integrative approaches (González et al., 2013). A Circos plot (Singh et al., 2019) can be
343 viewed as a generalization of relevance network where the nodes are located on a circle.
344 Then, based on the same pairwise similarity matrix used for relevance network, a clustered
345 image map can be displayed. This type of representation is based on a hierarchical clustering
346 simultaneously operating on the rows and columns of a real-valued similarity matrix. This is
347 graphically represented as a 2-dimensional colored image, where each entry of the matrix is
348 colored on the basis of its value, and where the rows and columns are reordered according to
349 the hierarchical clustering.

350

351 5. RESULTS

352 In this section, we provide neither a thorough biological interpretation of the results,
353 nor a comprehensive view of every statistical analysis performed. Instead, we highlight the
354 limits of each method leading to the next step of the statistical analysis and show how a
355 biologist can interpret and take over the conclusions of a statistical study.

356 5.1. Bivariate analysis

357 We illustrate the bivariate analysis through some graphical representations of
358 phenotypic data linked to one parameter of the experimental design. Figure 5A displays
359 parallel boxplots as well as individual observations of the number of leaves for the 5 ecotypes
360 at the 2 growth temperature conditions. Figure 5B only displays the average values of one
361 triplicate for each ecotype and temperature.

362 The main information extracted from these graphics concerns a quality control of the
363 data. The relatively low scattering of points representing individuals of each biological
364 replicate (Figure 5A) indicates a rather good reproducibility between all the samples and
365 between the repetitions. So, the values from several plants of a given biological repetition can
366 be averaged, to go on with the analyses. The visual impression provided by Figure 5B
367 regarding the temperature and ecotype effects can be confirmed via statistical testing such as
368 two-way ANOVA (Bingham & Fry) (results not shown). However, this kind of analysis does
369 not provide any information about the potential relationships between several variables. This
370 drawback justifies the next step of analysis which deals with a whole data set.

371

372 5.2. Multivariate analysis

373 The multivariate approach is illustrated on the rosettes cell wall transcriptomics data
374 set. It is composed of 364 variables (or transcripts). The first way to question the whole data
375 set can be through the computation of pairwise correlation coefficients. For instance, Figure 6

376 displays the correlation matrix between samples. It indicates that the levels of gene expression
377 for each sample are positively correlated (only green color and identically oriented ellipses)
378 with all the others.

379 Then, a PCA can be performed as an extension of the quality control. For instance,
380 Figure 7A highlights the distance between the three replicates corresponding to one condition.
381 We can observe that the Grip ecotype is well gathered, whereas the Col ecotype is more
382 scattered. This information must be moderated because of the rather low proportion of
383 variance explained by the first two principal components displayed here. Having a look at the
384 following components could be meaningful to consolidate and complete this information.

385 However, the interpretation of the PCA brings a first trend. Indeed, the samples are
386 clearly separated along the first (horizontal) axis according to the temperature: samples at
387 22°C are all located on the left (negative coordinates on PC1), whereas samples at 15°C are
388 on the right. This indicates that the effect of temperature is stronger than that of ecotypes
389 because PC1 capture the most important source of variability in the data. The representation
390 of the variables, *i.e.* the transcripts (Figure 7B), is not of great interest at this step; it mainly
391 highlights the need for selection methods to facilitate the interpretation of the results in terms
392 of gene expression level. However, the interpretation of such a plot jointly with the individual
393 plot enables, for instance, identifying over-expressed genes in samples at 15°C: they are
394 located on the right of the variables plot (in the same area as samples at 15°C in the individual
395 plot).

396

397 5.3. Supervised analysis and variable selection

398 To illustrate a supervised analysis, we deal with the same data set as before (cell wall
399 transcriptomics for the quantitative block) to discriminate the samples according to the

400 temperature (qualitative block) by performing a PLS-DA analysis. A similar analysis could be
401 made with the ecotype, but interpretation would be more complicated with 5 categories
402 instead of 2 for temperature. Moreover, we have already seen that the temperature effect is the
403 strongest for this data set (Figure 7A). Furthermore, to address the problem of interpretability
404 of the results, we also consider the sparse version of PLS-DA to select the most discriminant
405 genes for the temperature effect. The number of variables to select has to be determined by
406 the user. It depends on the way the potential candidates will be validated. For instance, if
407 validation has to be done through new biological experiments, the number of selected
408 variables must not be too large (about 10). But, if the validation consists in querying a
409 biological database, this number can be higher (about hundreds).

410 Figure 7 also displays the results of PLS-DA (C, D) and S-PLS-DA (E, F). Individuals
411 plots (Figure 7C, E) and variables plots (Figure 7D, F) are interpreted in the same way as
412 PCAs. Individuals plots only use two colors corresponding to the two temperatures. For both
413 PLS-DA and S-PLS-DA, the discrimination between the samples is clear-cut (Figure 7C, E).
414 This result confirms the overriding effect of the temperature. In other words, the variability
415 due to the five ecotypes does not impede from detecting the temperature effect. The result of
416 S-PLS-DA indicates that the discrimination can be observed with only a few genes. Indeed,
417 the difference between PLS-DA and S-PLS-DA relies on the number of genes involved in the
418 discrimination process. The list of the most relevant genes displayed in Figure 7F has to be
419 investigated through for instance functional analysis, but these developments are outside the
420 scope of this article.

421 These examples of sparse methods highlight the specificity of a supervised analysis: it
422 enables studying the impact of the factors of the experimental design (here the temperature)
423 on the quantitative variables. Thus, the biologist can play with these factors to answer its main
424 biological question and to identify potential future prospects.

425

426 5.4. Multi-block analyses

427 Multi-block analyses can address the main purpose of an integrative study by
428 analysing together all the blocks acquired for each sample. As an illustration, we expose the
429 results of a five-block supervised analysis focused on the rosettes, considering phenotypic,
430 cell wall transcriptomics, proteomics and metabolomics as quantitative variables and
431 temperature as the qualitative (or categorical) block.

432 The statistical relationships between blocks must be defined by the user through a
433 design matrix. This matrix is a square of size [(number of blocks) x (number of blocks)], it is
434 symmetrical and contains values between 0 and 1. A value close to or equal to 1 (respectively
435 0) indicates a strong relationship (respectively weak or no relationship) between the blocks to
436 be integrated. Fixing the values in the design matrix is crucial and complex because it requires
437 expressing biological relationships as numerical values (*e.g.* can we consider that the link
438 between proteomics and transcriptomics data is stronger than the link between proteomics and
439 metabolomics data?). For the sake of simplicity, 0 and 1 values can be used in a binary point
440 of view: blocks are linked or not. In a supervised context, the values also enable balancing the
441 optimisation between, on the one hand the relationships between quantitative blocks and, on
442 the other hand, the discrimination of the outcome. In our example, we considered a design
443 matrix composed of 0 between blocks to favor the discrimination task rather than the
444 relationships between the blocks. A full design matrix (composed of 1) highlights more
445 clearly relationships between blocks, but can lead to misclassified samples.

446 The interpretation of a multi-block supervised analysis requires several graphical
447 outputs. Some of them are presented in Figure 8. Figure 8A allows to check whether the
448 correlation between the first components from each data set has been maximized as specified
449 in the design matrix (Tenenhaus et al., 2014). Globally, correlation values are close to 1 and

450 mainly due to the separation of the two categories (22 vs 15°C; because of our design, this
451 matrix favors discrimination). With a full designed matrix, we get higher correlation values
452 but with less separated groups. Regarding the individual plots (Figure 8B), it appears that the
453 discrimination is better for the transcriptomics and proteomics blocks than for the others. The
454 sample plot (Figure 8B) has also to be interpreted regarding the variable plot (Figure 8C). To
455 make the interpretation easier, we present here the results of the sparse version of the multi-
456 block analysis. Therefore, we can identify variables from each block mainly involved in the
457 discrimination according to the temperature. For instance, variables located on the right on the
458 correlation circle plot (Figure 8C) contribute to the discrimination between the samples
459 growing at 22°C because they are also located on the right in the individuals plots (Figure
460 8B). Another way to display the results is presented in Figure 8D. The clustered image map
461 highlights the profiles of selected variables among the samples. It also includes the results of
462 hierarchical clustering performed jointly on variables and samples. Regarding the samples, the
463 two groups based on temperature are visualized through the dendrogram on the left. However,
464 let us note that the cluster gathering the samples at 15°C can be split into two sub-clusters
465 with the Col ecotype isolated. Regarding the variables, it mainly points out global trends of
466 the behavior of selected variables. The interpretation can then lead to retro analyses to
467 validate potential candidates. This can be done through new statistical analyses as well as new
468 biological experiments (Chawla et al., 2011).

469

470 5.5. Relevance networks

471 Another way to interpret the results of a multi-block approach consists in producing
472 relevance networks between variables. On Figure 9A, each selected variable is a node located
473 on a circle. Variables are sorted first according to their block, then depending on their

474 importance in discrimination. An edge links two nodes if their correlation is higher than a
475 threshold subjectively set by the user (we chose 0.9 in Figure 9A).

476 The correlations are mainly positive and concern a few variables from each block. To
477 complete the interpretation, we focus on another network generated with only two blocks
478 (Figure 9B, cell wall transcriptomics and proteomics). It accentuates the relationships between
479 pairs of proteins and transcripts. The selection of variables is a precious information for the
480 biologist to focus on some of them for validation and draw conclusions in biological terms.

481 Relevance networks can also be viewed as a first step to modelling as it mimics
482 biological networks and provides clues to address inference networks issues through further
483 dedicated experiments.

484

485 **6. CONCLUSION**

486 In an integrative biology context, the huge quantity of data produced, which can also
487 be heterogeneous, requires adapted and specific statistical methods tentatively summarized in
488 Figure 2. Even if the multi-block approaches can be viewed as the best tool to address a given
489 issue, other more basic standard statistical methods (univariate for instance) must not be
490 omitted. A deep understanding of a biological phenomenon requires a sequence of various
491 approaches to analyse the data. Finally, we consider that each method contributes to a better
492 interpretation of the others as we intended to express it with the schematic view of the
493 protocol as intertwined cycles (Figure 4). The statistical analysis of the large omics data sets
494 can be a never-ending story because each step of the framework provides information. The
495 results presented in this case study could not have been obtained using standard statistical
496 approaches. Actually, it is our global integrative strategy that led us to novel biological
497 results.

498

499 **Acknowledgements**

500 The authors are thankful to the Paul Sabatier-Toulouse 3 University and to the *Centre*
501 *National de la Recherche Scientifique* (CNRS) for granting their work. This work was also
502 supported by the French Laboratory of Excellence project "TULIP" (ANR-10-LABX-41;
503 ANR-11-IDEX-0002-02). HD was supported by the Midi-Pyrénées Region and the Federal
504 University of Toulouse. Thanks to Dr Kim-Anh Lê Cao, Pr Philippe Besse and François
505 Bartolo for their support and help with the graphical outputs and interpretation of the multi-
506 block analysis.

507

508 Supplementary Files:

509 Supplementary Table S1. Toy data set containing 12 observations and 3 variables.

510

511 **Figure captions:**

512 Figure 1. Scatterplot representing Vy values (vertical axis) according to Vx values (horizontal
513 axis). Control and treated observations are represented with grey triangles and black circles,
514 respectively.

515 Figure 2. Workflow for our multi-omics integrative studies. The different parts of this article
516 are represented with grey boxes and the green boxes close the workflow with biological
517 concepts. The workflow converges towards the functional analysis required to validate the
518 whole study.

519 Figure 3. Schematic overview of the strategy and experimental protocol used in this study.
520 Each circle represents one plant and each color stands for one ecotype of *A. thaliana*. For each
521 of the three biological replicates, the position of a given ecotype has been changed randomly
522 to avoid position effects.

523 Figure 4. One purpose, one method to analyse qualitative and quantitative blocks. A)
524 Schematic representation of the different blocks (or data sets) co-analysed in this study. The
525 samples are represented in rows and the variables in columns. B) Schematic overview of the
526 methods implemented represented by cycles within an integrative study. C) Examples of
527 graphical outputs detailed in the results section. PCA: Principal Component Analysis; MB:
528 Multi-Blocs; PLS: Partial Least Squares regression; DA: Discriminant Analysis. Qualitative
529 and quantitative blocks are represented in green and grey respectively.

530 Figure 5. Examples of graphical outputs of a supervised bivariate analysis illustrated by A) A
531 boxplot (each color corresponds to the different values obtained for each triplicate) and B) An
532 individual plot. (each color corresponds to the average obtained for one triplicate, and does
533 not match with color used in A). The number of leaves for 5 ecotypes of *A. thaliana* (Col,
534 Roch, Grip, Hern and Hosp) and 2 growth temperatures (22 and 15°C) was used. These plots

535 were obtained using functions `geom_point()` and `geom_boxplot()` from the `ggplot2` package
536 (Wickham, 2016).

537 Figure 6. A graphical representation of the multivariate analysis, pairwise correlation
538 coefficients of cell wall transcriptomics data sets in the rosettes of the five *A. thaliana*
539 ecotypes grown at 15°C or 22°C. The color code and the ellipse size represent the correlation
540 coefficient between the levels of expression of genes for each sample. The areas and the
541 orientations of the ellipses represent the absolute value of the corresponding correlation
542 coefficients. The eccentricity of the ellipses represents the absolute value of the corresponding
543 correlation coefficients. This plot was obtained using the function `corrplot()` from the `corrplot`
544 package (Wei & Simko, 2016).

545 Figure 7. Graphical representation of the unsupervised (A, B) and supervised (C-F) analysis
546 of the rosette cell wall transcriptomes from ecotypes grown at 22°C and 15°C. A) Individuals
547 plot of a PCA from ecotypes grown at 22°C (bright color) and 15°C (pale color) associated to
548 the B) Variables plot. C) Individuals plot of a PLS-DA from ecotypes grown at 22°C (orange)
549 and 15°C (blue) associated to the D) Variables plot and E) Individuals plot of a S-PLS-DA
550 associated to the E) Variables plot. Two circles of radius 1 and 0.5 are plotted in each
551 variables plot to reveal the correlation structure of the variables. These plots were obtained
552 using the functions `pca()`, `plsda()`, `plotIndiv()` and `plotVar()` from the `mixOmics` package
553 (Rohart et al., 2017).

554 Figure 8. A graphical representation of a multi-block analysis realised on the rosettes of
555 ecotypes grown at 22°C (orange) and 15°C (blue). A) `plotDIABLO` shows the correlation
556 between components from each data set maximized as specified in the design matrix. B)
557 Individuals plot projects each sample into the space spanned by the components of each block
558 associated to the C) Variables plot that highlights the contribution of each selected variable to
559 each component, D) Clustered image map of the variables (Protein: red; Transcripts: green;

560 Metabolites: grey; Phenotypes: black) to represent the multi-omics profiles for each sample
561 (15°C: blue, 22°C: orange). These plots were obtained using the functions `block.splsda()`,
562 `plotIndiv()`, `plotVar()` and `cim()` from the `mixOmics` package (Rohart et al., 2017).

563 Figure 9. Example of network representation. A) A Circos plot represents the correlations
564 between variables within and between each block (edges inside the circle) and shows the
565 average value of each variable in each condition (line profile outside the circle). B) A network
566 displaying the correlation between the transcriptomics (.T, green) and the proteomics data (.P,
567 red) colored from blue to red according to the color key. These plots were obtained using the
568 functions `circosPlot()` and `network()` from the package `mixOmics` (Rohart et al., 2017).

569

570 **References**

- 571 Anscombe, F. J. (1973). Graphs in statistical analysis. *Am. Stat.* 27(1), 17–21.
- 572 Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F.,
573 Huber, W. & Stegle, O. (2018). Multi Omics Factor Analysis—a framework for
574 unsupervised integration of multi omics data sets. *Mol. Syst. Biol.* 14:e8124 [https://](https://doi.org/10.15252/msb.20178124)
575 doi.org/10.15252/msb.20178124
- 576 Bingham, N. H., & Fry, J. M. (2010). *Regression: Linear models in statistics*: Springer
577 Science & Business Media.
- 578 Bécue-Bertaut, M., & Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of
579 quantitative, categorical and frequency data. *Comput. Stat. Data Anal*, 52(6), 3255-
580 3268.
- 581 Carey, M. A., & Papin, J. A. (2018). Ten simple rules for biologists learning to program.
582 *PLoS Comput Biol*, 14(1), e1005871. doi:10.1371/journal.pcbi.1005871
- 583 Chawla, K., Barah, P., Kuiper, M., & Bones, A. M. (2011). Systems biology: a promising tool
584 to study abiotic stress responses. *Omics and Plant Abiotic Stress Tolerance*, 163-172.
585 doi:10.2174/97816080505811110101
- 586 Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *J. Cell.*
587 *Biol.*, 177(1), 7-11. doi:10.1083/jcb.200611141
- 588 Duruflé, H., Hervé, V., Ranocha, P., Balliau, T., Zivy, M., Chourré, J., San Clemente, H.,
589 Burlat, V., Albenne, C., Déjean, S., Jamet, E., Dunand, C. (2017). Cell wall
590 modifications of two *Arabidopsis thaliana* ecotypes, Col and Sha, in response to sub-
591 optimal growth conditions: an integrative study. *Plant Sci.* 263, 183-193. doi:
592 10.1016/j.plantsci.2017.07. 015
- 593

- 594 Duruflé, H., Albenne, C., Jamet, E. & Dunand, C. Phenotyping and cell wall polysaccharide
595 composition of five *Arabidopsis* ecotypes grown at optimal or sub-optimal
596 temperatures, Data in brief (2019) (in press)
- 597 Duruflé, H., Ranocha, P., Balliau, T., Dunand, C. & Jamet, E. Transcriptomic and cell wall
598 proteomic datasets of rosettes and floral stems from five *Arabidopsis thaliana* ecotypes
599 grown at optimal or suboptimal temperature, Data in brief (2019) (in revision)
- 600 Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *Am. Stat.*,
601 56(4), 316-324.
- 602 Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Zhang,
603 J. (2004). Bioconductor: open software development for computational biology and
604 bioinformatics. *Genome Biol.*, 5(10), R80. doi:10.1186/gb-2004-5-10-r80
- 605 González, I., Lê Cao, K. A., Davis, M., & Déjean, S. (2013). Insightful graphical outputs to
606 explore relationships between two ‘omics’ data sets. *BioData Min.*, 5, 19.
- 607 Gray, S. B., & Brady, S. M. (2016). Plant developmental responses to climate change. *Dev.*
608 *Biol.*, 419(1), 64-77. doi:10.1016/j.ydbio.2016.07.023
- 609 Günther, O. P., Shin, H., Ng, R. T., McMaster, W. R., McManus, B. M., Keown, P. A., . . . Lê
610 Cao, K. A. (2014). Novel multivariate methods for integration of genomics and
611 proteomics data: applications in a kidney transplant rejection study. *OMICS*, 18(11),
612 682-695.
- 613 Hoffmann, M., H. (2002). Biogeography of *Arabidopsis thaliana* (L.) Heynh. (*Brassicaceae*).
614 *J. Biogeogr.*, 29(1), 125--134. doi:10.1046/j.1365-2699.2002.00647.x
- 615 Houben, K., Jolie, R., Fraeye, I., Van Loey, A., & Hendrickx, M. (2011). Comparative study
616 of the cell wall composition of broccoli, carrot, and tomato: Structural characterization
617 of the extractable pectins and hemicelluloses. *Carbohydr. Res.*, 346(9), 1105-1111.
618 doi:10.1016/j.carres.2011.04.014

- 619 Husson, F., & Josse, J. (2013). Handling missing values with/in multivariate data analysis
620 (principal component methods). *Agrocampus Ouest-Laboratoire de mathématique*
621 *appliquée, Rennes*.
- 622 Jamet, E., Roujol, D., San Clemente, H., Irshad, M., Soubigou-Taconnat, L., Renou, J. P., &
623 Pont-Lezica, R. (2009). Cell wall biogenesis of *Arabidopsis thaliana* elongating cells:
624 transcriptomics complements proteomics. *BMC Genomics*, *10*, 505. doi:10.1186/1471-
625 2164-10-505
- 626 Kerr, M. K. (2003). Experimental design to make the most of microarray studies. *Methods*
627 *Mol. Biol.*, *224*, 137-147. doi:10.1385/1-59259-364-X:137
- 628 Li, Y., Wu, F. X., & Ngom, A. (2016). A review on machine learning principles for multi-
629 view biological data integration. *Brief Bioinform.* doi:10.1093/bib/bbw113
- 630 Lê Cao, K. A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: biologically
631 relevant feature selection and graphical displays for multiclass problems. *BMC*
632 *Bioinformatics*, *12*, 253. doi:10.1186/1471-2105-12-253
- 633 Lê Cao, K. A., González, I., & Déjean, S. (2009). integrOmics: an R package to unravel
634 relationships between two omics datasets. *Bioinformatics*, *25*(21), 2855-2856.
635 doi:10.1093/bioinformatics/btp515
- 636 Maier, T., Güell, M., & Serrano, L. (2009). Correlation of mRNA and protein in complex
637 biological samples. *FEBS Lett*, *583*(24), 3966-3973. doi:10.1016/j.febslet.2009.10.036
- 638 Mardia, K. V., Kent, J. T., & Bibby, J. M. (1980). Multivariate analysis (probability and
639 mathematical statistics). Academic Press London.
- 640 Matejka, J., & Fitzmaurice, G. (2017). *Same stats, different graphs: Generating datasets with*
641 *varied appearance and identical statistics through simulated annealing*. Paper
642 presented at the Proceedings of the 2017 CHI Conference on Human Factors in
643 Computing Systems.

- 644 McDonald, J. H. (2009). *Handbook of biological statistics* (Vol. 2): Sparky House Publishing
645 Baltimore, MD.
- 646 Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., & Culhane, A. C.
647 (2016). Dimension reduction techniques for the integrative analysis of multi-omics
648 data. *Brief Bioinform.*, *17*(4), 628-641. doi:10.1093/bib/bbv108
- 649 Murdoch, D. J., & Chow, E. D. (1996). A graphical display of large correlation matrices. *The*
650 *Am. Stat.*, *50*(2), 178-180.
- 651 R Core Team, (2018). R: A Language and Environment for Statistical Computing.
- 652 Rai, A., Saito, K., & Yamazaki, M. (2017). Integrated omics analysis of specialized
653 metabolism in medicinal plants. *Plant J.*, *90*(4), 764-787. doi:10.1111/tpj.13485
- 654 Rajasundaram, D. Selbig, J. (2016). More effort — more results: recent advances in
655 integrative ‘omics’ data analysis, *Curr. Opin. Plant Biol.*, *30*, 57-61, [https://doi.org/](https://doi.org/10.1016/j.pbi.2015.12.010)
656 [10.1016/j.pbi.2015.12.010](https://doi.org/10.1016/j.pbi.2015.12.010).
- 657 Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for
658 ‘omics feature selection and multiple data integration. *PLoS Comput. Biol.*, *13*(11),
659 e1005752. doi:10.1371/journal.pcbi.1005752
- 660 Sabatier, R., Vivien, M., & Reynès, C. (2013). Une nouvelle proposition, l’analyse
661 discriminante multitableaux: Stasis-Ida. *Journal de la Société Française de Statistique*,
662 *154*(3), 31-43.
- 663 Savo, V., Lepofsky, D., Benner, J. P., Kohfeld, K. E., Bailey, J., & Lertzman, K. (2016).
664 Observations of climate change among subsistence-oriented communities around the
665 world. *Nat. Clim. Change*, *6*(5), 462-473.
- 666 Sibout, R. (2017). Crop breeding: Turning a lawn into a field. *Nat Plants*, *3*, 17060.
667 doi:10.1038/nplants.2017.60

- 668 Singh, A., Gautier, B., Shannon, C. P., Vacher, M., Rohart, F., Tebutt, S. J., & Le Cao, K. A.
669 (2019). DIABLO-an integrative, multi-omics, multivariate method for multi-group
670 classification. *Bioinformatics*, doi:10.1093/bioinformatics/bty1054.
- 671 Tenenhaus, A., Philippe, C., Guillemot, V., Lê Cao, K. A., Grill, J., & Frouin, V. (2014).
672 Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3),
673 569-583. doi:10.1093/biostatistics/kxu001
- 674 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser.*
675 *B-Stat. Methodol.*, 267-288.
- 676 Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., & González, I. (2016). Handling missing
677 rows in multi-omics data integration: multiple imputation in multiple factor analysis
678 framework. *BMC Bioinformatics*, 17(1), 402. doi:10.1186/s12859-016-1273-5
- 679 Wei, T., & Simko, V. (2016). corrplot: Visualization of a Correlation Matrix. R package
680 version 0.77. *CRAN, Vienna, Austria*.
- 681 Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1-23.
- 682 Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*: Springer.
- 683 Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics.
684 *Chemometrics Intell. Lab. Syst.*, 58(2), 109-130.
- 685 Zargar, S. M., Gupta, N., Nazir, M., Mir, R. A., Gupta, S. K., Agrawal, G. K., & Rakwal, R.
686 (2016). Omics—A New Approach to Sustainable Production. In *Breeding Oilseed*
687 *Crops for Sustainable Production* (pp. 317-344): Elsevier.

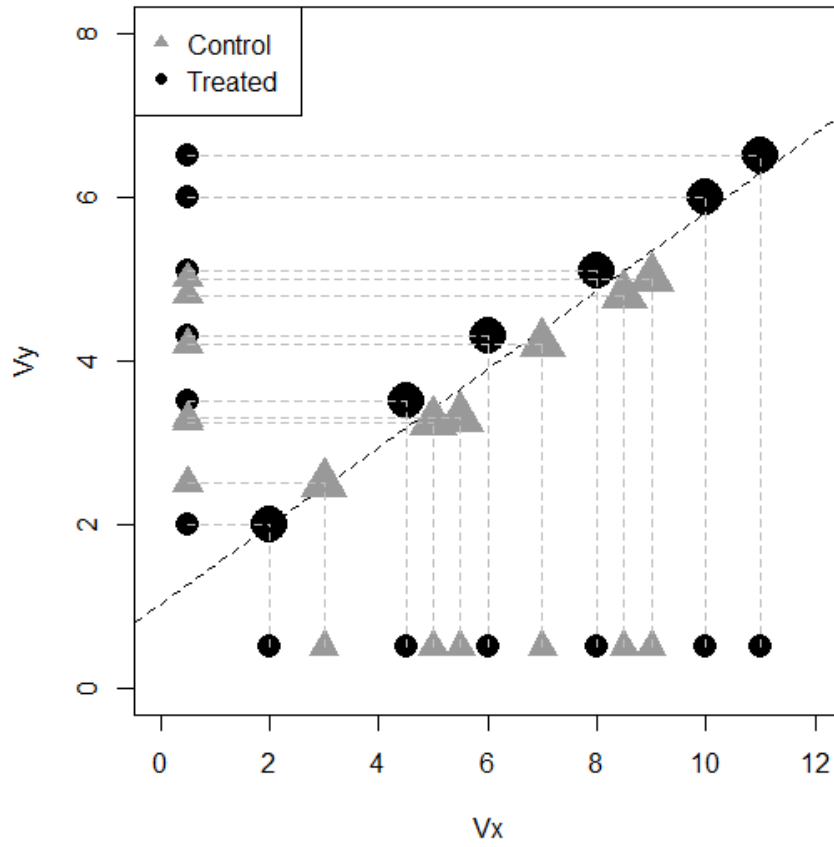


Figure 1. Scatterplot representing Vy values (vertical axis) according to Vx values (horizontal axis). Control and treated observations are represented with grey triangles and black circles, respectively.

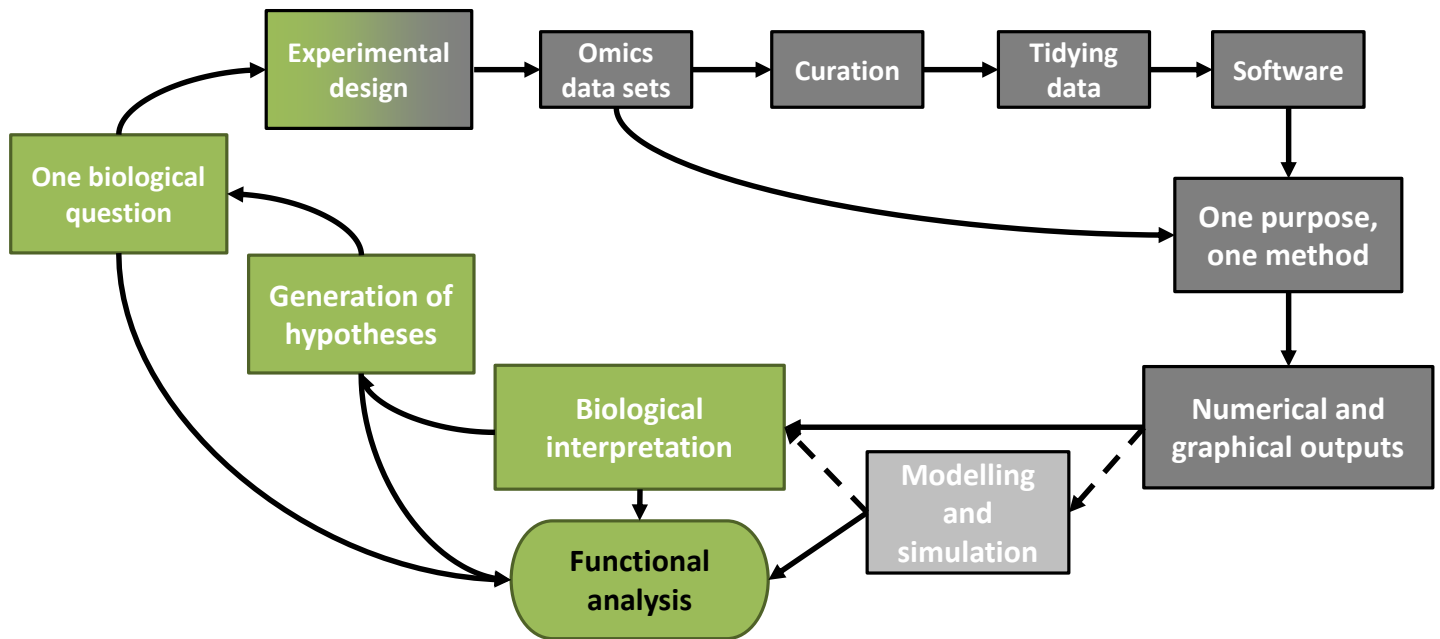


Figure 2. Workflow for our multi-omics integrative studies. The different parts of this article are represented with grey boxes and the green boxes close the workflow with biological concepts. The workflow converges towards the functional analysis required to validate the whole study.

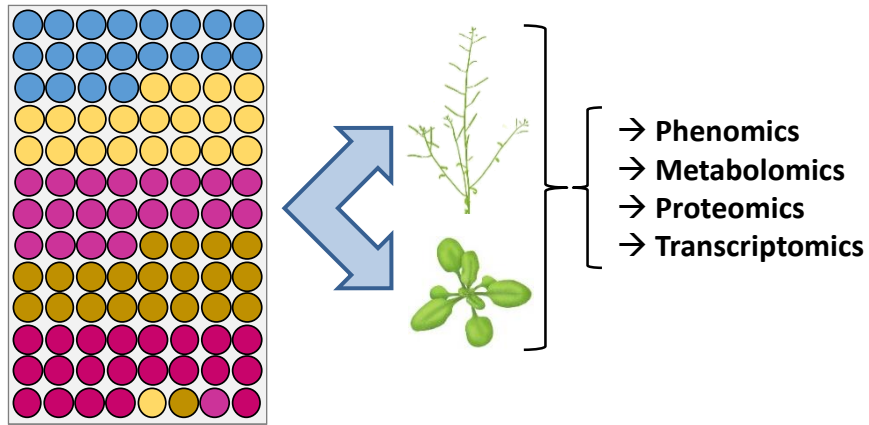


Figure 3. Schematic overview of the strategy and experimental protocol used in this study. Each circle represents one plant and each color stands for one ecotype of *A. thaliana*. For each of the three biological replicates, the position of a given ecotype has been changed randomly to avoid position effects.

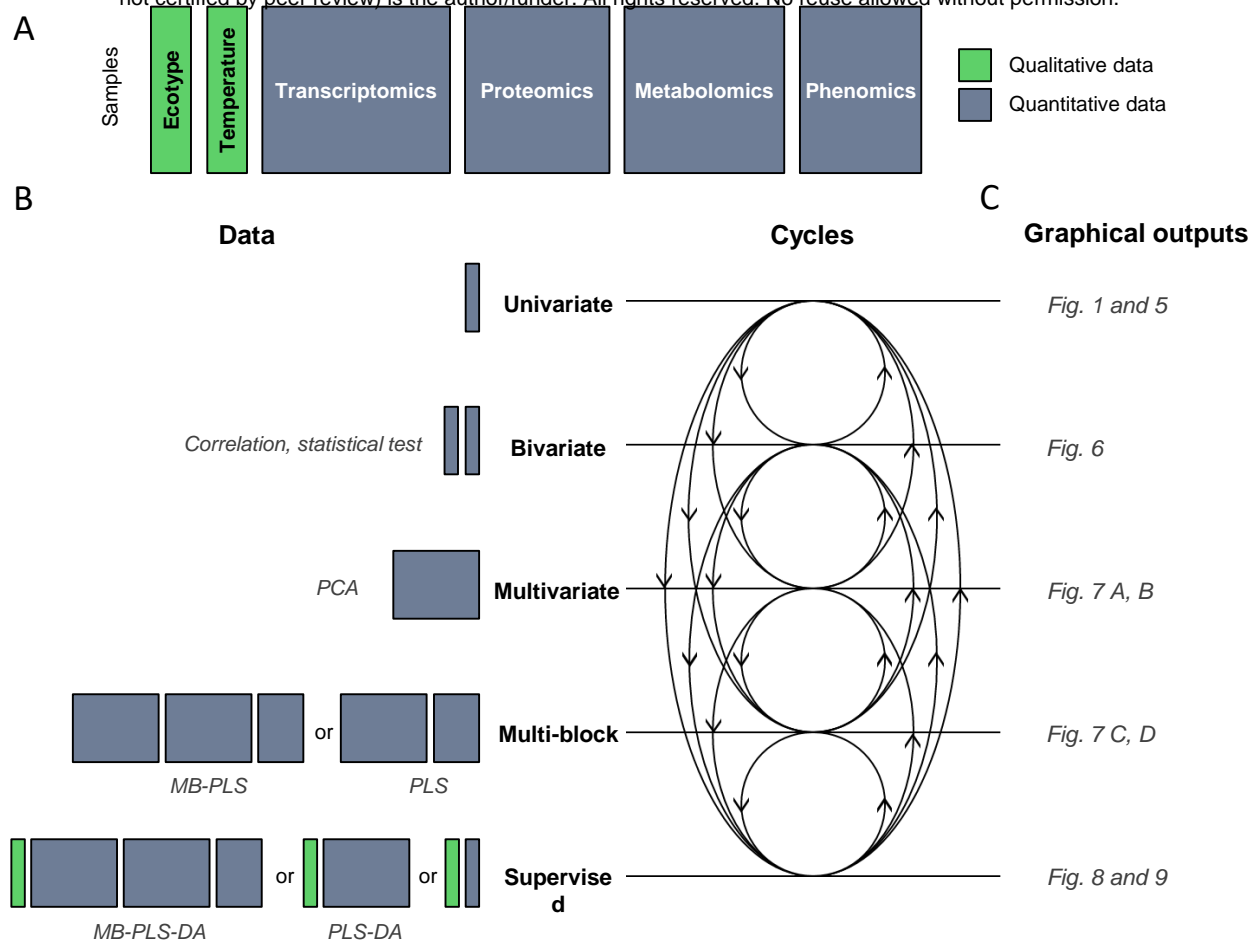
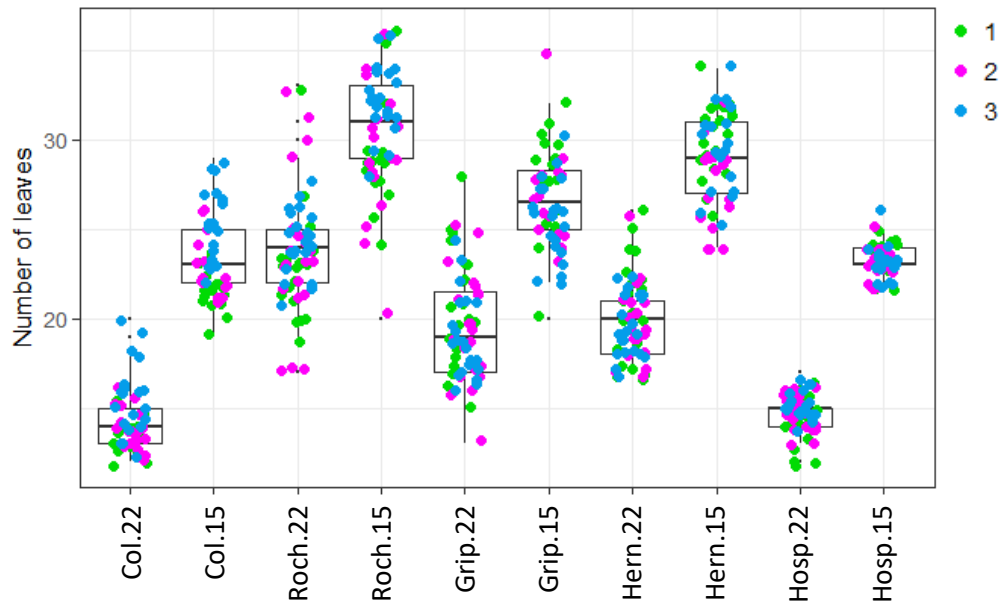


Figure 4. One purpose, one method to analyse qualitative and quantitative blocks. A) Schematic representation of the different blocks (or data sets) co-analysed in this study. The samples are represented in rows and the variables in columns. B) Schematic overview of the methods implemented represented by cycles within an integrative study. C) Examples of graphical outputs detailed in the results section. PCA: Principal Component Analysis; MB: Multi-Blocs; PLS: Partial Least Squares regression; DA: Discriminant Analysis. Qualitative and quantitative blocks are represented in green and grey respectively.

A



B

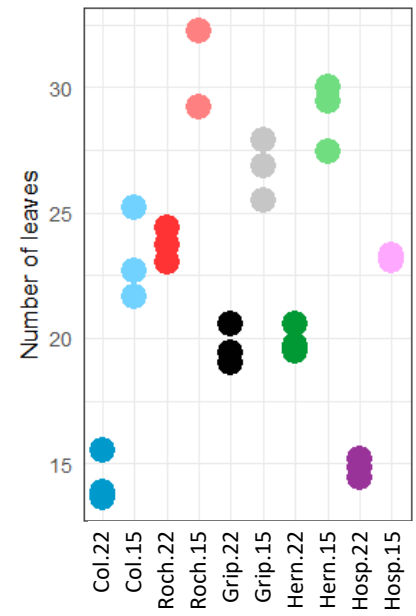


Figure 5. Examples of graphical outputs of a supervised bivariate analysis illustrated by A) A boxplot (each color corresponds to the different values obtained for each triplicate) and B) An individual plot. (each color corresponds to the average obtained for one triplicate, and does not match with color used in A). The number of leaves for 5 ecotypes of *A. thaliana* (Col, Roch, Grip, Hern and Hosp) and 2 growth temperatures (22 and 15°C) was used. These plots were obtained using functions `geom_point()` and `geom_boxplot()` from the `ggplot2` package (Wickham, 2016).

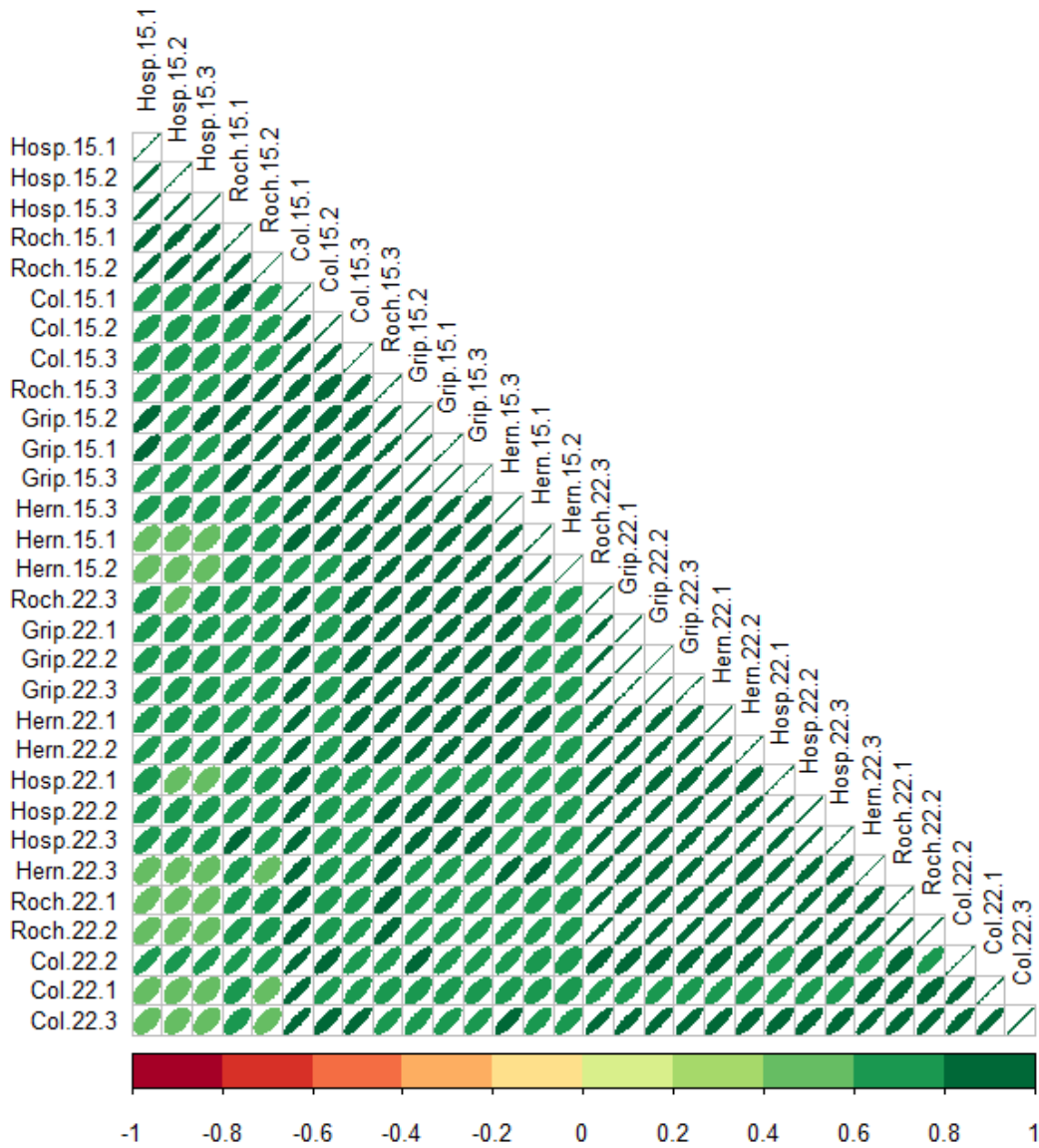


Figure 6. A graphical representation of the multivariate analysis, pairwise correlation coefficients of cell wall transcriptomics data sets in the rosettes of the five *A. thaliana* ecotypes grown at 15°C or 22°C. The color code and the ellipse size represent the correlation coefficient between the levels of expression of genes for each sample. The areas and the orientations of the ellipses represent the absolute value of the corresponding correlation coefficients. The eccentricity of the ellipses represents the absolute value of the corresponding correlation coefficients. This plot was obtained using the function `corrplot()` from the `corrplot` package (Wei & Simko, 2016).

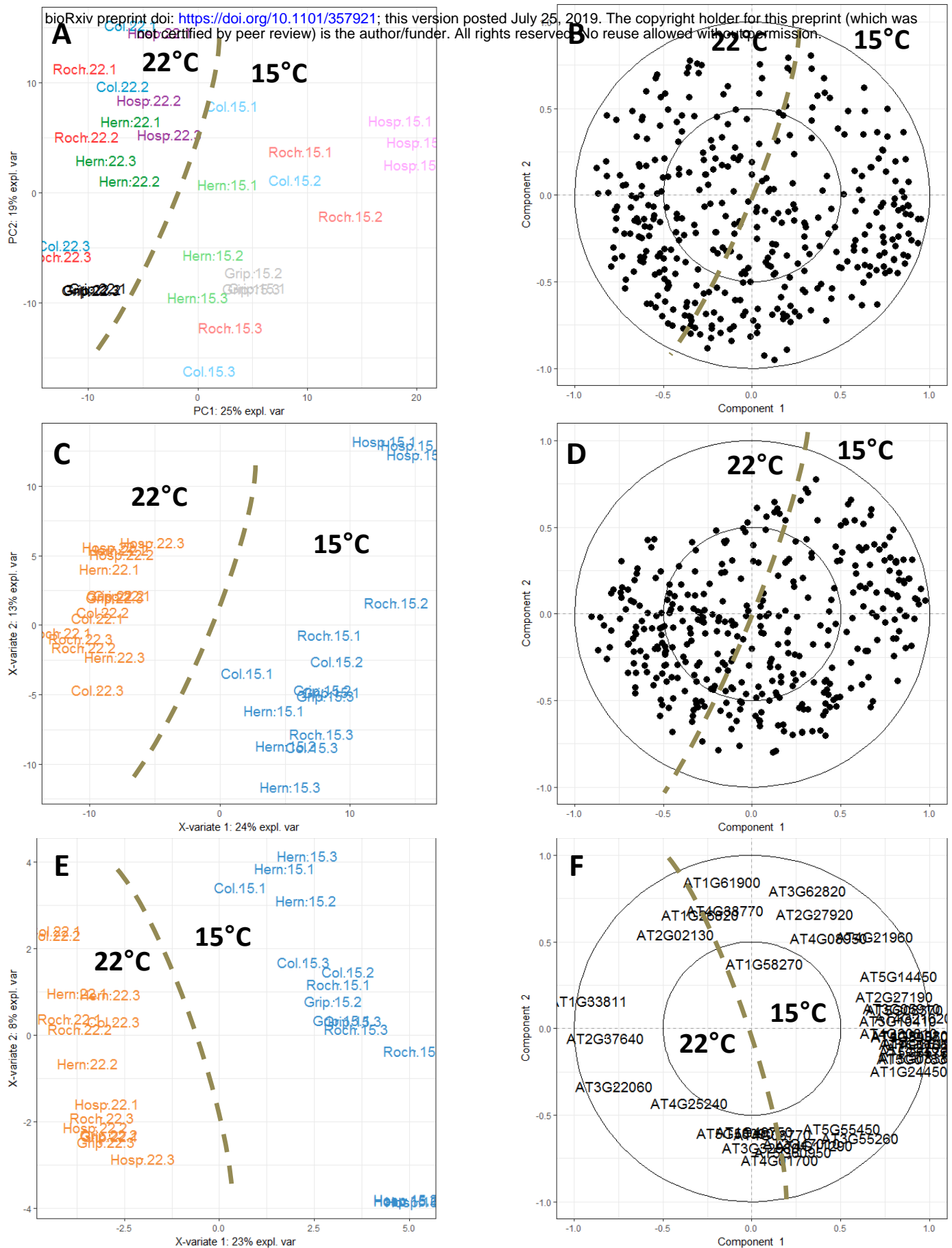


Figure 7. Graphical representation of the unsupervised (A, B) and supervised (C-F) analysis of the rosette cell wall transcriptomes from ecotypes grown at 22°C and 15°C. A) Individuals plot of a PCA from ecotypes grown at 22°C (bright color) and 15°C (pale color) associated to the B) Variables plot. C) Individuals plot of a PLS-DA from ecotypes grown at 22°C (orange) and 15°C (blue) associated to the D) Variables plot and E) Individuals plot of a S-PLS-DA associated to the E) Variables plot. Two circles of radius 1 and 0.5 are plotted in each variables plot to reveal the correlation structure of the variables. These plots were obtained using the functions `pca()`, `plsda()`, `plotIndiv()` and `plotVar()` from the `mixOmics` package (Rohart et al., 2017).

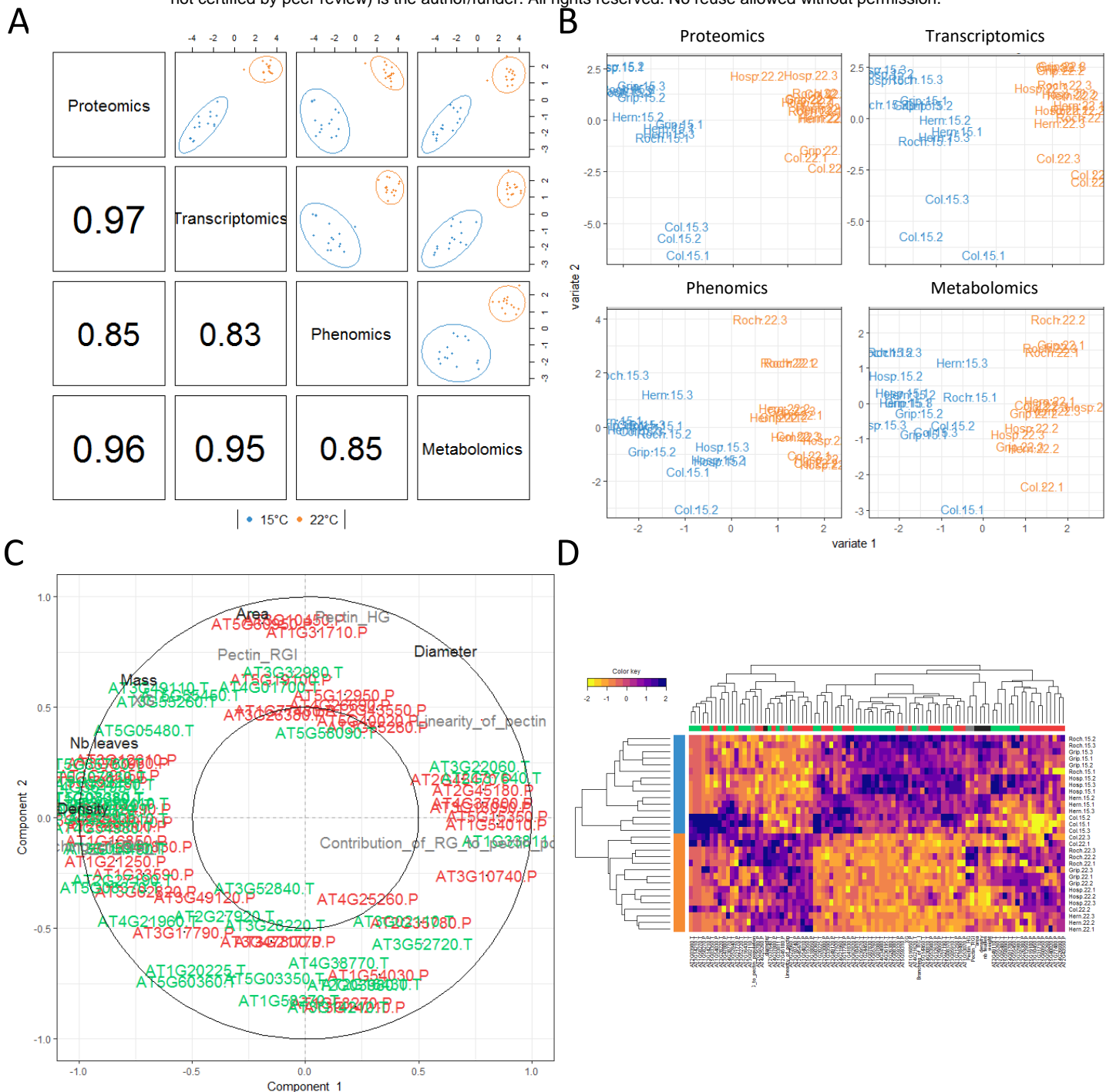
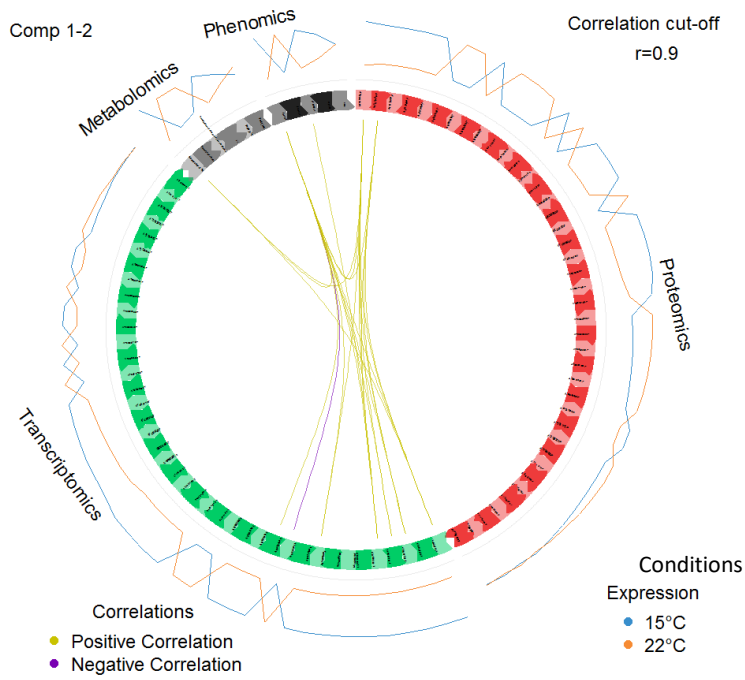


Figure 8. A graphical representation of a multi-block analysis realised on the rosettes of ecotypes grown at 22°C (orange) and 15°C (blue). A) plotDIABLO shows the correlation between components from each data set maximized as specified in the design matrix. B) Individuals plot projects each sample into the space spanned by the components of each block associated to the C) Variables plot that highlights the contribution of each selected variable to each component, D) Clustered image map of the variables (Protein: red; Transcripts: green; Metabolites: grey; Phenotypes: black) to represent the multi-omics profiles for each sample (15°C: blue, 22°C: orange). These plots were obtained using the functions `block.splsda()`, `plotIndiv()`, `plotVar()` and `cim()` from the `mixOmics` package (Rohart et al., 2017).

A



B

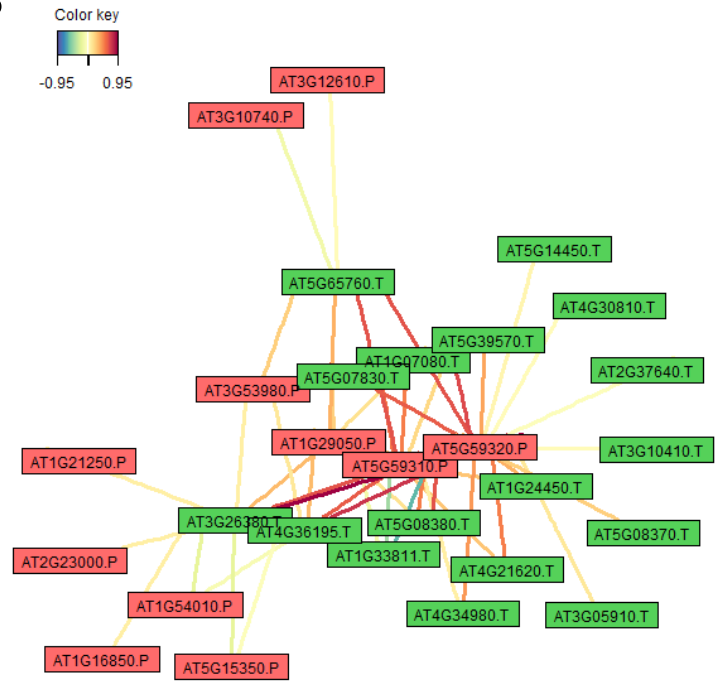


Figure 9. Figure 9. Example of network representation. A) A Circos plot represents the correlations between variables within and between each block (edges inside the circle) and shows the average value of each variable in each condition (line profile outside the circle). B) A network displaying the correlation between the transcriptomics (.T, green) and the proteomics data (.P, red) colored from blue to red according to the color key. These plots were obtained using the functions `circosPlot()` and `network()` from the package `mixOmics` (Rohart et al., 2017).