Fast computation of genome-metagenome interaction effects

Florent Guinot, Marie Szafranski, Julien Chiquet, Anouk Zancarini, Christine Le Signor, Christophe Mougel, Christophe Ambroise

FAST COMPUTATION OF GENOME-METAGENOME INTERACTION EFFECTS

TECHNICAL REPORT

Florent Guinot
L'Oréal, R&D, 94550, Chevilly-Larue, France
Université Paris-Saclay, CNRS, Univ Évry, Laboratoire de Mathématiques et Modélisation d'Évry
91037, Évry-Courcouronnes, France

Marie Szafranski
ENSIIE, 91025, Évry-Courcouronnes, France
Université Paris-Saclay, CNRS, Univ Évry, Laboratoire de Mathématiques et Modélisation d'Évry
91037, Évry-Courcouronnes, France
marie.szafranski@math.cnrs.fr

Julien Chiquet
Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75005, Paris, France

Anouk Zancarini
Plant Hormone Biology, Swammerdam Institute for Life Sciences, University of Amsterdam
1098 XH, Amsterdam, The Netherlands

Christine Le Signor
UMR 1347 Agroécologie, AgroSup Dijon, CNRS, Univ. Bourgogne, INRAE, Univ. Bourgogne Franche-Comté
21000, Dijon, France

Christophe Mougel
UMR 1349 IGEPP, INRAE, Agrocampus Ouest, Univ. Rennes 1, 35653, Le Rheu, France

Christophe Ambroise
Université Paris-Saclay, CNRS, Univ Évry, Laboratoire de Mathématiques et Modélisation d'Évry
91037, Évry-Courcouronnes, France

June 16, 2020

ABSTRACT

Motivation. Association studies have been widely used to search for associations between common genetic variants observations and a given phenotype. However, it is now generally accepted that genes and environment must be examined jointly when estimating phenotypic variance. In this work we consider two types of biological markers: genotypic markers, which characterize an observation in terms of inherited genetic information, and metagenomic marker which are related to the environment. Both types of markers are available in their millions and can be used to characterize any observation uniquely.

Objective. Our focus is on detecting interactions between groups of genetic and metagenomic markers in order to gain a better understanding of the complex relationship between environment and genome in the expression of a given phenotype.

Contributions. We propose a novel approach for efficiently detecting interactions between complementary datasets in a high-dimensional setting with a reduced computational cost. The method, named SICOMORE, reduces the dimension of the search space by selecting a subset of supervariables in the two complementary datasets. These supervariables are given by a weighted group structure defined on sets of variables at different scales. A Lasso selection is then applied on each type of supervariable to obtain a subset of potential interactions that will be explored via linear model testing.

Results. We compare SICOMORE with other approaches in simulations, with varying sample sizes, noise, and numbers of true interactions. SICOMORE exhibits convincing results in terms of recall, as well as competitive performances with respect to running time. The method is also used to detect interaction between genomic markers in *Medicago truncatula* and metagenomic markers in its rhizosphere bacterial community.

Software availability. A `R` package is available [Ambroise et al., 2020], along with its documentation and associated scripts, allowing the reader to reproduce the results presented in the paper.

***K*eywords** Statistical machine learning, variable selection, dimensionality reduction. Gene-environment interactions, genetic and metagenomic markers.

1 Introduction

Association studies are a popular approach for digging out genetic information relating to a given phenotype. To avoid confusion effects (e.g. stratification due to population origin) and improve the diagnostic, it is common practice to integrate environmental data in the analysis. These additional variables are generally few in number, of the order of ten.

In this paper we propose a generic method for taking thousands or even millions of environmental variables into consideration, with the aim of finding significant interactions between these variables and genetic markers. We illustrate the proposed algorithm on the genome of *Medicago truncatula (Fabaceae, Plantae)* and metagenomic markers in its rhizosphere bacterial community, but it could be applied in many other contexts.

1.1 Gene-environment interactions

Genome-Wide Association Studies (GWAS) look for genetic markers linked to a phenotype of interest. Typically, hundreds of thousands of single nucleotide polymorphisms (SNPs) are analyzed with a limited sample size using high-density genotyping arrays. GWAS are a powerful tool for investigating the genetic architecture of complex biological processes and have been successful in identifying hundreds of associated variants. However, they have been able to explain only a small proportion of the phenotypic variations expected from classical linkage analysis [Manolio et al., 2009].

Some of the missing heritability may be uncovered by taking into account correlations among variables and epistasis [Stanislas et al., 2017, and references therein]. Another way to understand and improve the knowledge of complex phenotypes is to look at gene-environment interactions. If the contributions of genes and environment to a phenotype are examined separately and interactions between them ignored, this can give incorrect estimates of how much phenotypic variance is attributable to genes alone, to environment alone, and to genes and environment jointly.

Gene-environment interactions are clearly of great interest in medical genetics and epidemiology [Clavel, 2007, Thomas, 2010] but also in plant research regarding environmental adaptation issues [Hancock et al., 2011, Hassani et al., 2018]. In particular, Metagenome-Wide Association Analysis (MWAS) [Segata et al., 2011, Wang et al., 2017, Wang and Jia, 2016] is providing a growing body of evidence regarding the role of gut microbiome in basic biological processes and in the development and progression of major human diseases, such as infectious diseases, gastrointestinal cancers, and metabolic diseases. In plants, the role of rhizosphere[1] microbiome on the plant growth and health is well known and has been studied since the early 2000s [Mukerji et al., 2002, Pinton et al., 2007, Lugtenberg and Kamilova, 2009, Berendsen et al., 2012]. While GWAS analyses have been able to identify associations between the plant genome of *Arabidopsis thaliana* and the metagenome (amplicon sequencing) of its associated phyllosphere and root microbial communities [Horton et al., 2014, Bergelson et al., 2019], in plants, to our knowledge, no specific MWAS analyses have so far been done.

[1]The rhizosphere was defined by Hiltner in 1904 as the area around a plant root that is inhabited by a unique population of microorganisms influenced by the chemicals released from plant roots.

1.2 Combining genome and metagenome analyses

There have been a number of works regarding the integration of multi-omics data in statistical or machine learning models, with several review papers. For instance, Li et al. [2016] establish a typology regarding different families of models. Huang et al. [2017] also list the kind of omics data which can be used and the outputs given by the methods. Hawe et al. [2019] pay attention to the inference of interaction networks.

However, these methods do not include environmental variables and consequently fail to address specificities of such features. There exists literature discussing both microbiome and genetics. They are mainly classical methods applied to a reduced set of species-gene pairs [Knights et al., 2014]. Another way of relating genetic and metagenomic data is to consider the metagenome as a phenotype and to perform quantitative trait locus (QTL) mapping. This kind of metagenomic QTL analysis illustrates the role of host genetics in shaping metagenomic diversity between individuals [Srinivas et al., 2013, Wang et al., 2016].

An alternative of interest is to consider metagenomic variables as environmental variables in GWAS. Several quantitative approaches have been proposed in classical gene-environment interaction studies with a small number of environmental factors limited to certain modalities, such as different status (smoking / non smoking, for instance) or medical treatments [Hutter et al., 2013, Han and Chatterjee, 2018]. More specifically, our proposal shares similarities with approaches where interactions can be modelled using a classical (generalized) linear model with interaction terms [Lin et al., 2013].

However, the number of interactions that need to be tested may increase dramatically when metagenomic markers are considered as environmental data. In this perspective, variable selection or variable compression may be of use here as a means of reducing the dimension of the problem in order to design an efficient method for detecting gene-environment interaction in a high-dimensional setting.

1.3 Taking structures into account in association studies

Data compression for dimension reduction may be achieved in various ways. A distinction is usually drawn between feature selection and feature extraction. Feature selection consists in selecting a few relevant variables from among the original variables, whereas feature extraction consists in computing new representative variables.

For the kind of association study that concerns us here, feature selection is often preferred to feature extraction for interpretative purposes. In this paper we advocate a mixed approach including feature extraction that is based on the underlying structures of genome and metagenome, combined with feature selection.

The idea of considering group structures is not new. It has already been advocated both in the context of GWAS [Dehman et al., 2015] and MWAS [Qin et al., 2012]. In the context of prediction from gene expression regression, Park et al. [2007] proposed clustering genes hierarchically to obtain a dendrogram that reveals their nested correlation structure. At each level of the hierarchy, supergenes are computed as the average expression of the current clusters. It can be shown that regressing over supergenes improves precision if the correlation structure is sufficiently strong. In a similar fashion, Guinot et al. [2018] made use of the haplotype structure of the human genome when they proposed a dimension-reduction approach that can be applied in the context of GWAS. It is worth noting that similar ideas have also been developed in other areas such medical imaging [Chevalier et al., 2018].

1.4 Contributions and organization of the paper

In this work, we propose a method for detecting interactions between genomic and metagenomic data. The method comprises four steps. Given a dataset:

(1) Identify a group structure within the variables using a hierarchical clustering;

(2) Create compressed features, or *supervariables*, according to this group structure;

(3) Select a subset of supervariables using a Lasso procedure with a penalty factor weighted by the length of the gap between two successive levels of a hierarchical clustering;

(4) Combine the two compressed datasets in a linear model with interactions in order to perform multiple hypothesis testing.

This scheme allows interactions to be detected efficiently in a high-dimensional setting with a reduced computational cost.

The paper is organized as follows. Section 2 looks at the role of linear models of interactions and proposes a framework for learning using complementary datasets. Section 3 describes our method, which seeks to uncover relevant interactions

using, first, compressions of data based on hierarchical structures, second, a Lasso selection procedure and, third, model testing. Finally, Section 4 provides an illustration of our approach using numerical simulations, and Section 5 describes an application for examining interactions between the genomic markers of the species *Medicago truncatula* and metagenomic markers of its rhizosphere microbial community.

2 Learning interactions with complementary datasets

This section gives a general introduction together with some notation, and outlines how we will establish a compact model of interactions between complementary datasets.

Remark. Here, and in what follows, the term *genomic* data will refer to SNP data. In Sections 2, 3 and 4, we will use the term *metagenomic* data for metabarcoding or shotgun data. The application on *Medicago truncatula* will be described in greater detail. Extensions to other kinds of data will be discussed in Section 6.

2.1 Setting and notations

Let us consider observations from two complementary views, G (for Genomic data) and M (for Metagenomic data), which are placed together in a training set $\mathcal{S} = \{(\mathbf{x}_i^G, \mathbf{x}_i^M, y_i)\}_{i=1}^N$, where $(\mathbf{x}_i^G, \mathbf{x}_i^M, y_i) \in \mathbb{R}^{D_G} \times \mathbb{R}^{D_M} \times \mathbb{R}$.

We assume the existence of underlying biological information on G and M, encoded as groups. The group structure over G is defined by N_G groups of variables $\mathcal{G} = \{\mathcal{G}_g\}_{g=1}^{N_G}$. We denote as $\mathbf{x}_i^g \in \mathbb{R}^{D_g}$ the sample i restricted to the variables of G from group $\mathcal{G}_g$. Similarly, the group structure over M is defined by N_M groups of variables $\mathcal{M} = \{\mathcal{M}_m\}_{m=1}^{N_M}$, and $\mathbf{x}_i^m \in \mathbb{R}^{D_m}$ is the sample i restricted to the variables of M from group $\mathcal{M}_m$.

We also introduce $D_I = D_G \cdot D_M$ and $N_I = N_G \cdot N_M$, corresponding to the number of variables and the number of groups that may interact.

Finally, we use the following convention: vectors of observations indexed with i, such as $\mathbf{x}_i$, will usually be row vectors, while vectors of coefficients, such as $\boldsymbol{\beta}$, will usually be column vectors.

2.2 Interactions in linear models

Interactions between data from views G and M may be captured in the model

$$y_i = \mathbf{x}_i^G \boldsymbol{\gamma}_G + \mathbf{x}_i^M \boldsymbol{\gamma}_M + \mathbf{x}_i^G \boldsymbol{\Delta}_{GM} (\mathbf{x}_i^M)^T + \epsilon_i \,, \tag{1}$$

where the vectors $\boldsymbol{\gamma}_G \in \mathbb{R}^{D_G}$ and $\boldsymbol{\gamma}_M \in \mathbb{R}^{D_M}$ denote the linear effects related to G and M respectively, the matrix $\boldsymbol{\Delta}_{GM} \in \mathbb{R}^{D_G \times D_M}$ contains the interactions between all pairs of variables in G and M, and $\epsilon_i \in \mathbb{R}$ is a residual error.

Models with interactions distinguish between *strong dependency* (SD) and *weak dependency* (WD). *Strong dependency* is the more common hypothesis (see for instance [Bien et al., 2013] and the discussion therein), and it means that an interaction is effective if and only if the corresponding single effects are also effective. *Weak dependency*, on the other hand, means that an interaction is effective if one of the main effects is also effective. Formally, for all variables $j \in \mathbf{x}^G$ and for all variables $j' \in \mathbf{x}^M$, if γ_j, $\gamma_{j'}$ and $\delta_{jj'}$ are the coefficients related to $\boldsymbol{\gamma}_G$, $\boldsymbol{\gamma}_M$ and $\boldsymbol{\Delta}_{GM}$, then

$$\begin{array}{lllll} (SD) & \delta_{jj'} \neq 0 & \Rightarrow & \gamma_j \neq 0 & \text{and} \quad \gamma_{j'} \neq 0 \,, \\ (WD) & \delta_{jj'} \neq 0 & \Rightarrow & \gamma_j \neq 0 & \text{or} \quad \gamma_{j'} \neq 0 \,. \end{array}$$

In this context, Bien et al. [2013] proposed a sparse model of interactions that is likely to encounter computational limitations for large-dimensional problems (Lim and Hastie [2015] and She et al. [2016]). Lim and Hastie [2015] present a method for learning pairwise interactions in a regression model by solving a constrained overlapping group Lasso [Jacob et al., 2009] in a manner that satisfies strong dependencies. She et al. [2016] propose a formulation with an overlapping regularization that fits both types of hypothesis, and they provide theoretical insights on the resulting estimators. [2]

However, the dimension $D_G + D_M + D_I$ inherent in Problem (1) when estimating $\boldsymbol{\gamma}_G$, $\boldsymbol{\gamma}_M$ and $\boldsymbol{\Delta}_{GM}$ may be inconveniently large, especially for applications with numerous variables such as in biology with genomic and metagenomic markers. To reduce this dimension we propose compressing the data according to an underlying structure that may be defined on the basis of prior knowledge or uncovered using clustering algorithms.

[2] To our knowledge, their implementation based on an alternating direction method of multipliers is not publicly available.

2.3 Compact model

Let us consider that if we have a compression function for all groups G and M, we can shape Problem (1) into a compact form

$$y_i = \sum_{g\in\mathcal{G}} \tilde{x}_i^g \beta_g + \sum_{m\in\mathcal{M}} \tilde{x}_i^m \beta_m + \sum_{g\in\mathcal{G}} \sum_{m\in\mathcal{M}} \underbrace{(\tilde{x}_i^g \cdot \tilde{x}_i^m)}_{\phi_i^{gm}} \theta_{gm} + \epsilon_i \,, \tag{2}$$

where $\tilde{x}_i^g \in \mathbb{R}$ is the i^{th} compressed sample of the variables that belong to the group g for the view G, and $\beta_g \in \mathbb{R}$ is its corresponding coefficient. The counterparts in the group m for the view M are $\tilde{x}_i^m \in \mathbb{R}$ and $\beta_m \in \mathbb{R}$. Finally, $\theta_{gm} \in \mathbb{R}$ is the interaction between groups g and m.

Problem (2) can be reformulated in a vector form. Let $\tilde{\mathbf{x}}_i \in \mathbb{R}^{N_G}$, $\boldsymbol{\beta}_G \in \mathbb{R}^{N_G}$, $\tilde{\mathbf{x}}_i \in \mathbb{R}^{N_M}$ and $\boldsymbol{\beta}_M \in \mathbb{R}^{N_M}$ be

$$\begin{aligned} \tilde{\mathbf{x}}_i^G &= (\tilde{x}_i^1 \cdots \tilde{x}_i^g \cdots \tilde{x}_i^{N_G}) \,, & \boldsymbol{\beta}_G &= (\beta_1 \cdots \beta_g \cdots \beta_{N_G})^T \,, \\ \tilde{\mathbf{x}}_i^M &= (\tilde{x}_i^1 \cdots \tilde{x}_i^m \cdots \tilde{x}_i^{N_M}) \,, & \boldsymbol{\beta}_M &= (\beta_1 \cdots \beta_m \cdots \beta_{N_M})^T \,. \end{aligned}$$

We denote as $\boldsymbol{\phi}_i \in \mathbb{R}^{N_I}$ the vector whose general component is given by ϕ_i^{gm} in Equation (2), that is

$$\boldsymbol{\phi}_i = \left(\phi_i^{11} \cdots \phi_i^{1N_M} \cdots \phi_i^{gm} \cdots \phi_i^{N_G 1} \cdots \phi_i^{N_G N_M}\right) \,,$$

and $\boldsymbol{\theta} \in \mathbb{R}^{N_I}$ denotes the corresponding vector of coefficients, that is

$$\boldsymbol{\theta} = (\theta_{11} \cdots \theta_{1N_M} \cdots \theta_{gm} \cdots \theta_{N_G 1} \cdots \theta_{N_G N_M})^T \,.$$

Finally, Problem (2) reads as a classical linear regression problem

$$y_i = \tilde{\mathbf{x}}_i^G \boldsymbol{\beta}_G + \tilde{\mathbf{x}}_i^M \boldsymbol{\beta}_M + \boldsymbol{\phi}_i \boldsymbol{\theta} + \epsilon_i \,, \tag{3}$$

of dimension $N_G + N_M + N_I$.

2.4 Uncovering relevant interactions

Compared to Problem (1) and provided that N_G and N_M are reasonably smaller than D_G and D_M, the dimension of Problem (3) is drastically reduced, so that it may be solved with the aid of a suitable optimization algorithm and sufficient computing resources. For instance, Donoho and Tsaig [2008] give an overview of ℓ_1 regularized algorithms to solve sparse problems like Lasso, which in our case could take the form:

$$\begin{cases} \underset{\boldsymbol{\beta}_G, \boldsymbol{\beta}_M, \boldsymbol{\theta}}{\operatorname{argmin}} & \displaystyle\sum_{i=1}^{n} \left(y_i - \tilde{\mathbf{x}}_i^G \boldsymbol{\beta}_G - \tilde{\mathbf{x}}_i^M \boldsymbol{\beta}_M - \boldsymbol{\phi}_i \boldsymbol{\theta}\right)^2 \\ & \quad + \lambda_G \displaystyle\sum_{g=1}^{N_G} |\beta_g| + \lambda_M \sum_{m=1}^{N_M} |\beta_m| + \lambda_I \sum_{g,m=1}^{N_I} |\theta_{gm}| \,, \end{cases}$$

with λ_G, λ_M and λ_I being the positive hyperparameters that respectively control the amount of sparsity related to coefficients $\boldsymbol{\beta}_G$, $\boldsymbol{\beta}_M$ and $\boldsymbol{\theta}$. The $N_G + N_M + N_I$ dimension may nevertheless remain large in relation to the number of observations N. Also, it will be remarked that this kind of formulation does not automatically entail the dependency hypotheses (SD) and (WD) unless additional constraints are introduced. For this purpose, the works by Bien et al. [2013], Lim and Hastie [2015] or She et al. [2016] mentioned above may be considered. In the following section we present another way of reducing the dimension further and ensuring that the strong dependency hypothesis is satisfied.

3 Method

In this section we provide some elements for addressing Problem (3) in relation to biological problems involving complementary datasets. Our proposed approach, which we have named **SICOMORE** (Selection of Interaction effects in COmpressed Multiple Omics REpresentations), is available for download as an `R` package [Ambroise et al., 2020].

3.1 Preprocessing of the data

When tackling problems that involve genomic and metagenomic interactions, some prior transformations are necessary. This preliminary step may also include a first attempt at reducing the dimension.

Transformation for metagenomic data

Metagenome sequencing gives rise to features that take the form of proportions in different samples. This kind of information is referred to in the statistical literature as compositional data [Aitchison, 1982] and is known to be subject to negative correlation bias [Pearson, 1896, Aitchison, 1982]. The most common way to circumvent this issue is to transform the D_M features using centered log-ratios and to replace 0 values using maximum-likelihood approaches (see [Gloor et al., 2016, 2017] and references therein). A more detailed presentation of these aspects may be found in [Rau, 2017].

Initial selection of variables

As described in Section 2, we make the assumption that interactions have strong dependencies, which means that an interaction can be effective only if the two simple effects associated with the variables in interaction are included in the model. For this reason it may be advantageous to make an initial selection in order to eliminate inoperative single effects on G and M respectively. Different approaches for carrying out this selection may be considered. For example, screening rules can eliminate variables that will not contribute to the optimal solution of a sparse problem, sweeping all the variables upstream to the optimization. In cases where this kind of screening is appropriate, the work of Lee et al. [2017] is a useful resource. Their focus is on Lasso problems and they present an overview of these techniques, together with an ensemble of screening rules. Once the screening has been performed, the optimization of a Lasso problem gives the final set of variables.

3.2 Structuring the data

Once the data have been preprocessed, hierarchical clustering using Ward's method with appropriate distances can be employed to uncover the tree structures.

Clustering of metagenomic data

Several approaches are available for analyzing microbiota compositions. Li [2015] has produced a review of statistical and computational methods according to different objectives and/or technologies. For problems with numerous similar reference sequences, Fischer et al. [2017] have proposed a general linear model approach designed to estimate taxon abundances for strain-level analyses.

A commonly used approach when analyzing metabarcoding data is to group sequences into taxonomic units [Blaxter et al., 2005]. The features arising from such a sequencing are often modeled as Operational Taxonomic Units (OTUs), each OTU representing species proxies according to some degree of sequence similarity. More recent methods based on denoising techniques have led to the definition of Amplicon Sequence Variants (ASVs), which can be considered as refined versions of OTUs [Callahan et al., 2017].

While the structure of microbial communities can be defined according to the underlying phylogenetic tree, it also makes sense to use more classical distances to define a hierarchy based on the abundance of OTUs. In our application, we use an agglomerative hierarchical clustering with the Ward criterion.

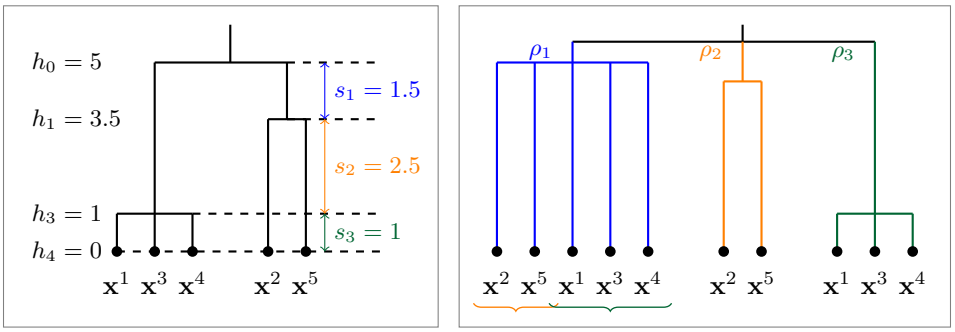
Clustering of genomic data

When the genomic information is available through SNP, the tree structure on G will be defined using a hierarchical clustering algorithm that integrates the linkage disequilibrium as the measure of dissimilarity [Dehman et al., 2015].

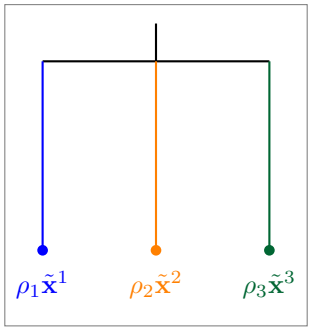
This algorithm is a computationally efficient hierarchical clustering that makes use of the structure of the genome in order to cluster SNPs into adjacent groups. More specifically, it is a spatially constrained hierarchical clustering based on Ward's incremental sum-of-squares algorithm [Ward, 1963] in which the measure of dissimilarity is based on the linkage disequilibrium between SNPs j and j': $1 - r^2(j, j')$. The algorithm also makes use of the fact that the linkage disequilibrium matrix can be modeled as block-diagonal by allowing only groups of variables that are adjacent on the genome to be merged, which significantly reduces the computational cost.

3.3 Using the structure efficiently

Different approaches for finding an optimal number of clusters may be envisaged when looking for the optimal cut in a tree structure obtained by hierarchical clustering (see for instance [Milligan and Cooper, 1985] or [Gordon, 1999]). Whatever the approach, finding this optimal cut necessarily involves a systematic exploration of different levels of the hierarchy. Our alternative strategy for bypassing this expensive exploration is as follows:



(a) Original hierarchy (b) Expanded representation



(c) Compressed representation

Figure 1: Dimension reduction strategy. (a) Original hierarchical tree with an example for 5 variables. (b) Expanded representation of the tree with all possible weighted groups derived from the original hierarchy. The group in blue gathers the variables contained in the groups in orange and green. (c) Compressed representation of the tree after construction of the supervariables.

(a) Expanding the hierarchy, considering all possible groups at a single level;

(b) Assigning a weight to each group based on the distances between two consecutive groups in the hierarchy;

(c) Compressing each group into a supervariable.

The different steps in this strategy are illustrated in Figure 1, from the original tree structure in Figure 1(a) to the final flattened, weighted, compressed representation shown in Figure 1(c).

Expanding the hierarchy **(a)**

To reduce the dimension of Problem (3), the first step consists in flattening the respective tree structures obtained on views G and M so that only one group structure remains. Each group of variables defined at the deepest level may thus be included in other groups of larger scales, as shown in Figure 1(b).

Assigning weights to the groups **(b)**

To keep track of the tree structure, an additional measure may be included to quantify the loss of information between two successive levels. More specifically, for a tree structure of height H and for $1 \leq h \leq H-1$, we define s_h as the gap between heights h and $h-1$. Using a similar methodology to Grimonprez [2016] for the multi-layer group Lasso, we define this quantity as $\rho_h = 1/\sqrt{s_h}$. The process is shown in Figure 1(a) and 1(b).

Compressing the data **(c)**

To summarize each group of variables the mean, the median, or other quantiles may be used, as well as more sophisticated representations based on eigenvalue decomposition, such as the first factor of a Principal Component Analysis.

3.4 Identification of relevant supervariables

With the aid of this compressed representation we can uncover relevant interactions using a multiple testing strategy.

Selection of supervariables

Compression is a key ingredient in reducing significantly the dimension of Problem (3). We take this a step further with an additional feature selection process applied to the compressed variables, as described at the beginning of this section, in order to preprocess the data using screening rules and/or applying a Lasso optimization on each view G and M:

$$\underset{\boldsymbol{\beta}_G}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \tilde{\mathbf{x}}_i^G \boldsymbol{\beta}_G\right)^2 + \lambda_G \sum_{g=1}^{N_G} \rho_g |\beta_g| \,,$$

and

$$\underset{\boldsymbol{\beta}_M}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \tilde{\mathbf{x}}_i^M \boldsymbol{\beta}_M\right)^2 + \lambda_M \sum_{m=1}^{N_M} \rho_m |\beta_m| \,,$$

with penalty factors defined by $\rho_g = 1/\sqrt{s_g}$ and $\rho_m = 1/\sqrt{s_m}$, as explained in Section 3.2.

This step for selecting the supervariables in the two complementary datasets can be subject to instability when setting the amount of selection. The method can be improved further in terms of model consistency by using resampling techniques [Bach, 2008, Meinshausen and Bühlmann, 2010, Hofner et al., 2015]. This has been implemented in SICOMORE with the `R` package stabs [Benjamin and Hothorn, 2017].

Linear model testing

For the purpose of feature selection the relevant interactions may be uncovered separately by considering each selected group $g \in \mathcal{G}$ coupled with each selected group $m \in \mathcal{M}$ in a linear model of interaction and by performing a hypothesis test (a standard t-test for instance) on each parameter θ_{gm}:

$$y_i = \tilde{x}_i^g \beta_g + \tilde{x}_i^m \beta_m + (\tilde{x}_i^g \cdot \tilde{x}_i^m)\, \theta_{gm} + \epsilon_i \,. \tag{4}$$

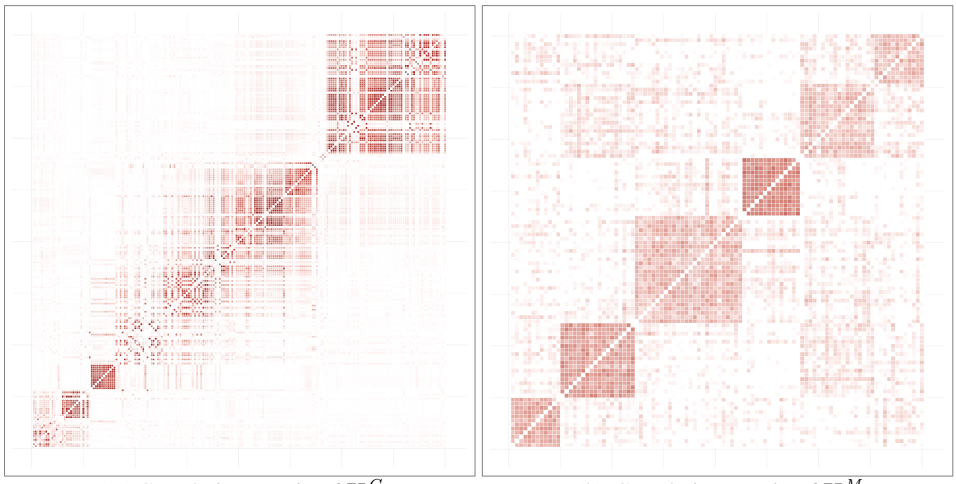
This strategy has the advantage of highlighting all the potential interactions between the selected simple effects in an exploratory rather than a predictive analysis perspective. It can also be seen as an alternative way of shortcutting Problem (3), in that it involves N_I problems of dimension 3 rather than a potentially large problem of dimension $N_G + N_M + N_I$. Finally, by construction, this selection scheme preserves strong dependencies.

4 Numerical simulations

We present some numerical simulations to assess SICOMORE's ability to uncover relevant interactions. We compare our approach with two other methods, namely **MLGL** [Grimonprez, 2016] and **glinternet** [Lim and Hastie, 2015]. These two methods will be described in more detail later in the section. Both are available as `R` packages on the CRAN platform [Grimonprez et al., 2020, Lim and Hastie, 2019].

These numerical simulations are designed to study several aspects of SICOMORE:

- The ability to recover relevant interactions will be observed on different configurations with respect to the sample sizes, the noise, and the number of true interactions.
- The impact of the weighting scheme will be shown with two versions of our approach, using both weighted and unweighted supervariables.
- The impact of the compression scheme will be compared to MLGL using the same structure but with the initial variables.
- Finally, a dedicated simulation sketches the running times necessary for each method to reach convergence when the dimension of one of the matrices grows. To allow the comparison of SICOMORE with MLGL or glinternet, the dimensions of the simulated matrices have been kept between a few hundred and a few thousand.



(a) Correlation matrix of $\mathbf{X}^G$ (b) Correlation matrix of $\mathbf{X}^M$

Figure 2: Examples of group structures: correlations observed on (a) genomic data $\mathbf{X}^G$ and (b) metagenomic data $\mathbf{X}^M$.

4.1 Data generation

Generation of metagenomic and genomic data matrices

Genomic data. To obtain a matrix $\mathbf{X}^G$ resembling real genomic data we used `HAPGEN2` software [Su et al., 2011a,b], which can simulate an entire chromosome conditionally on a reference set of population haplotypes (from HapMap3) and an estimate of the fine-scale recombination rate across the region, so that the simulated data share similar patterns with the reference data. We generated chromosome 1 using the haplotype structure of CEU population (Utah residents with Northern and Western European ancestry from the CEPH[3] collection) as the reference set, and we selected $D_G = 200$ variables from this matrix to obtain the simulated dataset. An example of the linkage disequilibrium structure among the simulated SNPs is shown in Figure 2(a).

Metagenomic data. The data matrix $\mathbf{X}^M$, with $D_M = 100$ variables, was generated using a multivariate Poisson log-normal distribution [Aitchison and Ho, 1989] with group structure dependencies. The Poisson log-normal model is a latent Gaussian model where latent vectors $\mathrm{Z}_i \in \mathbb{R}^{D_M}$ are drawn from a multivariate normal distribution

$$\mathrm{Z}_i \sim \mathcal{N}_{D_M}(0, \mathbf{\Sigma}) \,,$$

and where $\mathbf{\Sigma}$ is a covariance matrix that can give a correlation structure between the variables. The random variable X_i^M related to the centered phenotypic count data is then drawn from a Poisson distribution conditionally on Z_i

$$\mathrm{X}_{ij}^M | \mathrm{Z}_{ij} \sim \mathcal{P}\left(e^{\mu_j + \mathrm{Z}_{ij}}\right).$$

The group structure shown in Figure 2(b) was obtained by drawing a latent multivariate normal vector using a covariance matrix such that the correlation level between the latent variables in a group are between 0.5 and 0.95. Simulating in this way gives a matrix of count data with a covariance structure close to what is observed with metagenomic data. As described in Section 3.1, we computed the proportions for each of the random variables and transformed them using centered log-ratios.

Generation of the phenotype

For all simulations we used a fixed value of $N_M = 6$ groups for the matrix $\mathbf{X}^M$. For the matrix $\mathbf{X}^G$, since `HAPGEN2` does not allow the group structure to be controlled exactly, we used the gap statistic [Tibshirani et al., 2001] to identify a number of groups in the hierarchy. For instance, in Figure 2(a), the gap statistic identified $N_G = 16$ groups. The supervariables were then calculated using averaged groups of variables to obtain the two matrices of supervariables, $\tilde{\mathbf{X}}^G$ and $\tilde{\mathbf{X}}^M$.

To generate the phenotype, we considered a data structure for which the data to regress was generated using supervariables according a linear model with interactions of the form:

[3] `http://www.cephb.fr`.

$$y_i = \sum_{g \in \mathcal{S}^G} \tilde{x}_i^g \beta_g + \sum_{m \in \mathcal{S}^M} \tilde{x}_i^m \beta_m + \sum_{g \in \mathcal{S}^G} \sum_{m \in \mathcal{S}^M} \underbrace{(\tilde{x}_i^g \cdot \tilde{x}_i^m)}_{\phi_i^{gm}} \theta_{gm} + \epsilon_i \,, \tag{5}$$

where $\mathcal{S}^G$ and $\mathcal{S}^M$ are subsets of randomly chosen effects from the matrices $\tilde{\mathbf{X}}^G$ and $\tilde{\mathbf{X}}^M$ respectively, $\tilde{x}_i^g$ is the i^{th} sample of the g effect and β_g its corresponding coefficient, and $\tilde{x}_i^m$ is the i^{th} sample of the m effect and β_m its corresponding coefficient. Finally, θ_{gm} is the interaction between variables $\tilde{x}_i^g$ and $\tilde{x}_i^m$.

We considered $I \in \{1, 3, 5, 7, 10\}$ true interactions between some supervariables to generate the phenotype such that I blocks of the coefficients of θ_{gm} have non zero values. The process was repeated 30 times for each couple of parameters in $N = \{50, 100, 200\} \times sd(\boldsymbol{\epsilon}) = \{0.5, 1, 2\}$.

4.2 Comparison of methods

In accordance with the outline given in the preamble of Section 4, we were seeking to assess the ability of SICOMORE, in comparison with MLGL and glinternet, to uncover true causal interactions. For this purpose, we needed to reshape the datasets provided to the two methods as we now describe below.

It is worth mentioning that SICOMORE is an approach that draws on the work of Park et al. [2007] and MLGL [Grimonprez, 2016], with an explicit design for detecting interactions. We explore two settings : ρ-SICOMORE and SICOMORE, which correspond respectively to the method described in section 3 using $\rho_h = 1/\sqrt{s_h}$ and $\rho_h = 1, \forall h$.

Multi-Layer Group Lasso **(MLGL)**

Grimonprez [2016] defines MLGL as a two-step procedure that combines a hierarchical clustering with a group Lasso regression. It is a weighted version of the overlapping group Lasso [Jacob et al., 2009] which performs variable selection on multiple group partitions defined by the hierarchical clustering. A weight is attributed to each possible group identified at all levels of the hierarchy, as described in Section 3(b). This weighting scheme favors the creation of groups associated with large gaps in the hierarchy.

The model of interactions is fitted with weights on the groups defined by the expanded representation of the two hierarchies using the initial variables, as illustrated in Figure 1(b). The ability of MLGL to uncover real interactions is evaluated positively if it selects the correct interaction terms between two groups of variables at the right level in both hierarchies.

It should be noted that here MLGL is not being evaluated in a context for which it was intended, since MLGL examines the different levels of a hierarchical structure using *all* variables. This approach is not well suited in a high-dimensional setting and still less in a model of interactions. But, as we explained at the beginning of Section 4, this comparison with MLGL is intended to shed light on the impact of the compression applied to the variables in SICOMORE.

Group Lasso interaction network **(glinternet)**

Lim and Hastie [2015] introduced glinternet, a procedure that considers pairwise interactions in a linear model in a way that satisfies strong dependencies between main and interaction effects: whenever an interaction is estimated to be non-zero, its two corresponding main effects are also included in the model.

It fits a hierarchical group Lasso model, with constraints on the main and interactions effects, as specified in Section 2.4, and it accommodates the strong dependency hypothesis by adding an appropriate penalty to the loss function (we refer the reader to [Lim and Hastie, 2015] for more details on the form of the penalty). For very large problems (with a number of variables $\geq 10^5$), the group Lasso procedure is preceded by a screening step that gives a candidate set of main effects and interactions.

Since this method can only work at the level of variables, we needed to include a group structure into the analysis, and so we decided to fit the glinternet model on the compressed variables and to constrain the model to only fit the interaction terms between the supervariables of the two matrices $\tilde{\mathbf{X}}^G$ and $\tilde{\mathbf{X}}^M$. We explicitly removed all interaction terms between supervariables belonging to the same data matrix.

To ensure that our comparison of SICOMORE was fair, we considered two options, namely GLtree and GLgap. The GLtree option works on the unweighted compressed representations of the two hierarchies (Figure 1(c)) and thus takes into account all the possible interactions between the supervariables of the two datasets. In contrast, the GLgap option considers only the interactions between the compressed variables constructed at a specific level in the hierarchies, chosen by the gap statistic. Given that D^G and D^M are the numbers of variables in $\mathbf{X}^G$ and $\mathbf{X}^M$, the dimension

of the matrices $\tilde{\mathbf{X}}^G$ and $\tilde{\mathbf{X}}^M$ in GLtree are respectively $\tilde{D}^G = D^G + (D^G - 1)$ and $\tilde{D}^M = D^M + (D^M - 1)$. [4] Consequently, for GLtree the number of interactions to be examined is $\tilde{D}^G \times \tilde{D}^M$, while for GLgap this number will depend on the level chosen by the gap statistic, but it will necessarily be smaller since this option considers only a specific level of the hierarchy. In the numerical simulations, given that $D^G = 200$ and $D^M = 100$, the use of strong rules to discard variables is therefore not necessary.

4.3 Evaluation metrics

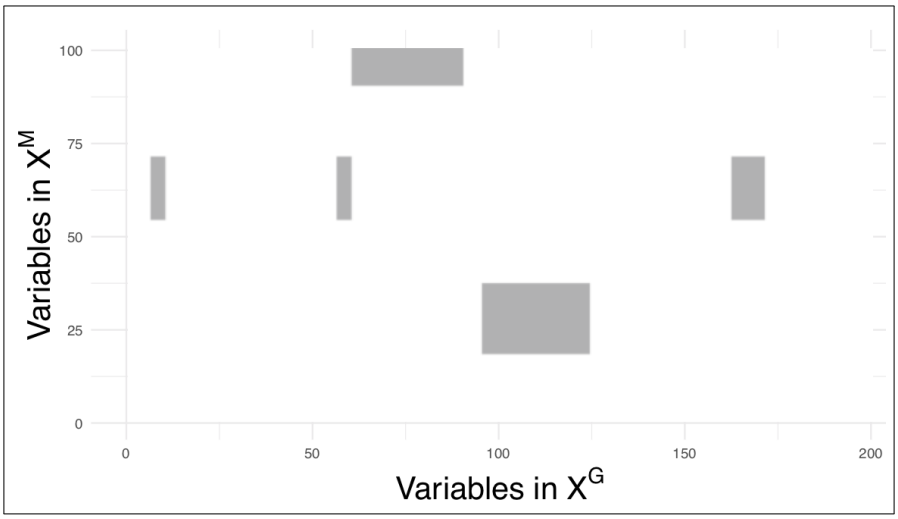


Figure 3: **Illustration of the true interaction matrix $\boldsymbol{\theta}$** with $I = 5$, $\sigma = 0.5$ and $n = 100$. Each non-zero value in this matrix is considered as a true interaction between two variables.

For each run we evaluated the quality of the variable selection using Precision and Recall. More precisely, we compared the true interaction matrix $\boldsymbol{\theta}$ that we used to generate the phenotype with the estimated interaction matrix $\hat{\boldsymbol{\theta}}$ computed for each model.

For all possible $D_G \times D_M$ interactions, with $\theta_{jj'}$ the interaction term between variable $j \in \mathbf{X}^G$ and variable $j' \in \mathbf{X}^M$, we determined the following confusion matrix:

	$\hat{\theta}_{jj'} = 0$	$\hat{\theta}_{jj'} \neq 0$
$\theta_{jj'} = 0$	True Negative	False Positive
$\theta_{jj'} \neq 0$	False Negative	True Positive

The performances are measured with Precision $= \frac{TP}{FP+TP}$ and Recall $= \frac{TP}{FN+TP}$. An example of the interaction matrix $\hat{\boldsymbol{\theta}}$ is shown in Figure 3 for $I = 5$ blocks in interaction.

Here, a *true positive* corresponds to a significant p-value on a true causal interaction, a *false positive* to a significant p-value on a noise interaction, and a *false negative* to a non-significant p-value on a true causal interaction.

For the three tested methods we corrected for multiple testing by controlling the family-wise error rate with the Holm-Bonferroni method. Even though it is known to be stringent, we chose the Holm-Bonferroni method to adjust for multiple testing because the number of hypothesis tests that needed to be performed for our simulation was quite low. In a high-dimensional context, for example in analyzing real microarray data, the Benjamini-Hochberg method would be preferable for controlling the false discovery rate.

[4] In GLtree, a matrix $\tilde{\mathbf{X}}$ is created using the initial D variables, and the $(D-1)$ groups of variables of the dendogram from the hierarchical clustering are added as compressed features.

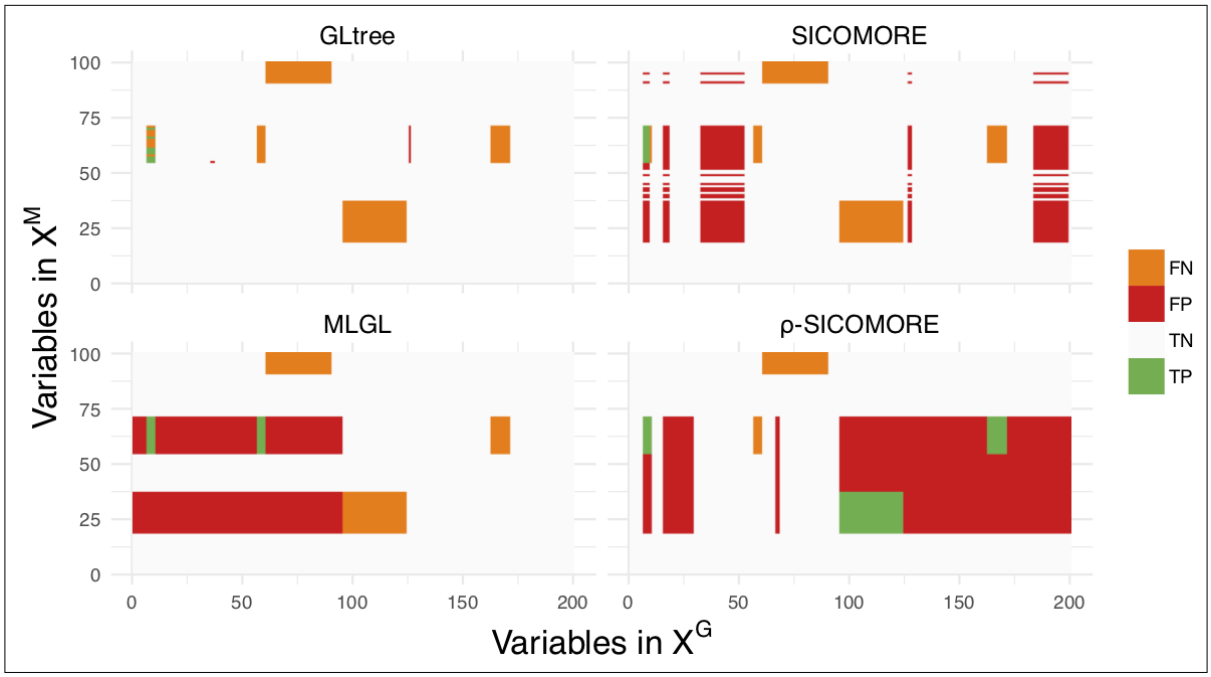


Figure 4: **Confusion matrices of interactions** $\hat{\theta}_{jj'}$ for the different methods, using the following simulation parameters: $I = 5$, $\sigma = 0.5$, $n = 100$. We can see from this example that MLGL and ρ-SICOMORE behave similarly, with very large genomic regions identified. SICOMORE tends to work with smaller genomic and metagenomic regions.

4.4 Performance results

The performances of the different methods in uncovering true causal interactions are shown in figures 5(a) (for Precision) and 5(b) (for Recall). For the sake of clarity we show only the results for $I = 7$ blocks of variables in interaction. The results for $I \in \{1, 3, 5, 10\}$ are provided in Appendix A as supplementary results. The plots in Figure 4 represent the uncovered confusion matrices of interaction θ_{gm} corresponding to one particular set of simulation parameters ($I = 5$, $\sigma = 0.5$, $n = 100$) for each of the compared methods.

The Recall results show that MLGL and ρ-SICOMORE are good at uncovering true positive interactions, with ρ-SICOMORE performing better overall. SICOMORE performs less well because it favours the selection of small groups that are only partly contained in the groups that generate the interactions. This indicates that MLGL and ρ-SICOMORE have an effective weighting scheme. GLgap is unable to uncover relevant interactions, but here the way the structure between variables is defined using the gap statistic differs from the other methods. The Precision results show that all methods perform poorly, with a significant number of false positive interactions. MLGL and ρ-SICOMORE tend to select groups of variables and supervariables that are too high in the tree structure, giving rise to false positives that are spatially close to the true interactions. SICOMORE, which, as explained above, favours small groups, gives fewer false positives of this kind. The behaviour of GLgap may vary according to the selected cut with the gap statistic into the tree structure, while the GLtree option has slightly better precision. Note that this improved precision may be the consequence of the additional information provided from our group definition. The glinternet method is mostly unable to uncover the true interactions correctly, whether the compressed or the original representation is used.

4.5 Computation time

In order to reduce the computation time required to run our algorithm, we chose to restrict the search space. It is limited to the area of the tree where the jumps in the hierarchy are the largest, and the number of groups to be evaluated is arbitrarily set to five times the number of initial features. This reduces the number of variables to be fitted in the Lasso regression but does not affect performance regarding Recall and Precision.

We compared the computational performance of our method with the two others by varying the number of variables in $\tilde{\mathbf{X}}^G$. We repeated the number of evaluation five times for each size of $\tilde{\mathbf{X}}^G$ and averaged the computation time.

N_G	50	100	500	1000	1500	2000	3000	4000
ρ-SICOMORE	0.01	0.01	0.02	0.03	0.03	0.04	0.05	0.06
SICOMORE	0.21	0.34	0.82	0.76	0.75	0.96	0.93	1.09
MLGL	0.06	0.09	3.35	0.86	3.12	4.52	8.02	24.20
GLtree	0.07	0.28	0.67	3.83	11.69	26.31	88.17	210.64

Table 1: Average computation time (in minutes) over 5 replicates for varying dimensions of $\tilde{\mathbf{X}}^G$, with the dimension of $\tilde{\mathbf{X}}^G$ being fixed ($N_M = 6$).

We can conclude from the results presented in Table 1 that two methods, glinternet and MLGL, are unsuitable for large-scale analyses of genomic data, since computation time starts to rise steeply once the number of variables exceeds a few thousand. The computation time of ρ-SICOMORE and SICOMORE is drastically reduced compared to MLGL or glinternet, with ρ-SICOMORE having a slight advantage due to the weighting scheme that induces faster elimination of non relevant supervariables.

5 Application on the rhizosphere bacterial communities of *Medicago truncatula*

For an implementation of our algorithm on real data we chose to study the interactions between the genome of *Medicago truncatula* and the metagenome (16S rRNA gene sequencing) of its rhizosphere bacterial community. We were seeking to identify significant interactions in order to better understand the effect of both the plant genome and the rhizosphere bacterial microbial community on plant growth.

For this purpose, a core collection of 155 accessions (all from INRAE-Montpellier) were grown in a controlled environment and phenotyped for several traits related to the plant growth and nutritional strategy:

- Total Dry Biomass (TDB).
- Root Total Dry Biomass Ratio (RTDBR).
- Specific Nitrogen Uptake (SNU) expressed as mg of $N.g^{-1}$ of belowground biomass per day.

In addition to the phenotypic measurement, the rhizosphere of each accession was also analyzed to determine the bacterial diversity and composition (see Appendix B). The metabarcoding raw data is available in the European Nucleotide Archive (ENA) EMBL-EBI database system under project accession PRJEB25849.

A total of 15617 different bacterial OTUs were found in the rhizosphere of the plants. The different OTUs were pooled according to their taxonomic affiliation at the genus level, and a total of 329 genera were thus analyzed.

The 155 sequenced accessions, extracted from `http://www.medicagohapmap.org`, were genotyped with a DNA microarray chip, giving a total of 6 372 968 SNPs after 3% MAF, multiallele SNP exclusion and minimum count (100) filtering. The missing values were imputed using the `snp.imputation` function from the `R` package snpStats [Clayton, 2019]. Given two sets of SNPs typed in the same subjects, this function computes rules that can be used to impute one set from the other in a subsequent sample. By discarding any SNP that had too many missing values to be completely imputed, we reduced the size of the data to 2 148 505 SNPs.

The positions of SNPs inside or in the vicinity of genes ($\pm$ 2Kb) were extracted from context files downloaded from `http://www.medicagohapmap.org`. A Singular Enrichment Analysis was conducted using an exact Fisher test with the `R` package topGO [Alexa and Rahnenfuhrer, 2019] and GO term annotation from `http://www.medicagogenome.org`.

The algorithm requires several hyper-parameters to be chosen in order to run properly:

- **Aggregating function**: For the genomic and the metagenomic data, we defined the mean value of the group as supervariable.
- **Clustering algorithm**: For the metagenomic data we used a hierarchical clustering using Ward's distance as the measure of similarity. For the genomic data we used a spatially constrained hierarchical clustering algorithm that integrates the linkage disequilibrium as the measure of dissimilarity.
- **Stability selection**: The parameters of the function stabs in SICOMORE for the metagenomic data were fixed to $\texttt{B} = 300$ subsampling replicates, with the frequency of selection of the supervariables on the replicates $\texttt{cutoff} = 0.7$. The upper bound for the per-family error rate was set to $\texttt{PFER} = 1$. For the genomic data, the parameters were fixed to $\texttt{B} = 100$, $\texttt{cutoff} = 0.6$ and $\texttt{PFER} = 10$.

- **Search space**: For computational reasons we chose to run some analyses chromosome by chromosome. Correction for multiple testing was done by controlling the false discovery rate [Benjamini and Hochberg, 1995]. Since weak effects were expected, we also examined interactions with p-values < 0.05 to discuss some aspects in relation with the phenotypes RTDBR and SNU.

Regarding the running time for the application, for about 2M SNPs and 329 bacterial genera, the algorithm was able to perform the analysis in 250 min ($\sim$ 4 hours) with 10 CPU cores (Intel(R) Xeon(R) CPU E7-480 @ 2.40GHz) and 2.5 Gb of memory.

Results regarding Total Dry Biomass

No significant interactions were found for this phenotype.

Results regarding the Root Total Dry Biomass Ratio

For RTDBR, four interactions were significant at p-value < 0.05, distributed across three chromosomes, as shown in Table 2. The 365 210 SNPs allow recovering 9 007 genes. A Gene Ontology enrichment analysis carried on the 4 490 annotated genes identified "hormone biosynthetic process" (Fisher p-value of 2.10^{-17}) or "antibiotic biosynthetic process" (Fisher p-value of 5.10^{-18}), "systemic acquired resistance" (Fisher p-value of 2.10^{-9}) and "cellular response to nitrogen starvation" (Fisher p-value of 2.10^{-8}) as four main overrepresented metabolic pathways involved in RTDBR variations under microbe interactions. The three first classes included almost redundant genes, mainly NBS-LRR kinase and 8 transcription factors. The fourth term "cellular response to nitrogen starvation" is composed mainly of lectin-domain receptor kinases genes also present in the three other classes and related to plant defense and of cysteine-rich receptor kinase genes, which are known to be regulated upon biotic and abiotic stress, such as salt and drought stress. For the rhizosphere bacterial communities, 39 genera were found in interaction with these genes. Also, 17, 9, and 6 genera were affiliated to Proteobacteria, Actinobacteria, and Bacteroidtes respectively. Within Proteobacteria, 10 genera were identified as Alphaproteobacteria and 4 of them to the *Rhizobiales* family, which is known to contribute to N nutrition of *Medicago truncatula*. Plant disease resistance genes play a major role in the plant immune system that was induced during pathogenic plant-microbial interactions but also during mutualistic plant-microbe interactions [Hacquard et al., 2017]. None of the 39 bacterial genera identified was affiliated to genera known as plant pathogens. However, several of the bacterial genera identified were affiliated to genera known as plant symbiont or plant growth promoting bacteria. We could hypothesize that bacteria affiliated to these genera could be in positive interaction with the plant and induced some defense response.

Results regarding Specific Nitrogen Uptake

For the SNU, we retrieved 157 698 significant SNPs and 5 476 genes from the three significant interactions, as shown in Table 2. Among the 3 136 annotated genes, the most over-represented biological process was the "transmembrane receptor protein tyrosine kinase signalling pathway" (Fisher p-value of 1.10^{-6}), "regulation of anion channel activity" (Fisher p-value of 3.10^{-4}) and "lignin biosynthesis" (Fisher p-value of 8.10^{-4}). The two first classes were partly redundant and mainly composed of LRR receptor kinase genes, known to be involved in plant innate immunity. The term "regulation of anion channel activity" was linked to other significant terms related to regulation to ion/anion transport. The "lignin biosynthesis" process included genes involved in lignin biosynthesis such as 8 caffeic acid O-methyltransferase genes, 3 cinnamyl alcohol dehydrogenase-like protein or 2 shikimate O-hydroxycinnamoyltransferase, which serve as building blocks in the formation of plant lignin [Tu et al., 2010]. The colonization of plant host cells by bacteria involves the progressive remodeling of the plant–microbial interface for both *Rhizobium*-Legume symbiosis [Brewin, 2004] and pathogen bacteria [Underwood, 2012]. In addition, the plant immune system is involved in symbiosis and during plant pathogen infections, and more generally with the plant microbiota [Gourion et al., 2015, Hacquard et al., 2017]. For the rhizosphere bacterial communities, 180 genera were found in interaction with these genes. 83, 31, 24 and 23 genera were affiliated to Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes respectively. In addition to the 13 genera belonging to the *Rhizobiales* family, other OTUs were affiliated to bacteria genera harboring functional traits relating to the N cycle, such as nitrogen fixation, nitrate reduction to ammonium, and denitrification, which can contribute to plant nitrogen nutrition.

Altogether, the mathematical method proposed here could support some biological hypothesis that need to be validated using other biological approaches combining plant mutant affected by these genes and simplified bacteria community defined on the genera identified.

PH	#MG	CHR	GP	#SNPs	p-value	q-value
RTDBR	39 genera	3	129:980206	6705	0.03	0.18
RTDBR	39 genera	3	980235:32366703	196705	0.04	0.18
RTDBR	39 genera	7	21704918:33495621	68658	0.03	0.23
RTDBR	39 genera	8	50:18024047	93142	0.02	0.14
SNU	180 genera	2	38539843:45729381	33033	0.04	0.13
SNU	180 genera	6	33985403:35275305	6174	0.04	0.13
SNU	180 genera	8	18024755:45569421	156827	0.05	0.09

Table 2: Results of the search for interactions using the ρ-SICOMORE method. From left to right, the names of the columns are: PH for the phenotype studied; #MG for the number of genera; CHR for the chromosome; GP for the genomic postion (pb) and #SNPs for the number of SNPs in the genomic region.

6 Conclusion

Synthesis

The detection of interaction effects in a high-dimensional setting remains a difficult problem because multiple testing is onerous and because effects are small in terms of their significance. In this work, we proposed SICOMORE, a method that reduces the dimension of the search space by selecting a subset of compressed variables obtained from the biological characteristics of complementary datasets.

Our approach has demonstrated its ability to uncover interaction effects with a high statistical power. In our simulations, where sample sizes, noise, and the number of true interactions all varied, SICOMORE always exhibited stronger recall than both MLGL and glinternet. SICOMORE combines the strengths of different methods in a powerful single algorithm. SICOMORE is also significantly more efficient than the others in terms of computation time.

SICOMORE was able to detect interactions between the genome of *Medicago truncatula* and its rhizosphere, which are linked to the Root Total Biomass Ratio as well as its Specific Nitrogen Uptake.

Extensions

Although our approach as presented here concerns the detection of interactions between genomic and metagenomic markers, it should be noted that two major extensions are available.

1. SICOMORE can be applied to any kind of numerical data, as long as an underlying hierarchical or group structure is available (such as a correlation structure, for instance). In particular, our method can handle shotgun sequencing as well as other omics data, or even clinical follow-up, which often takes the form of categorical data that can be easily structured.
2. The compression scheme used in SICOMORE means that the model of interactions can easily be extended to $V > 2$ different datasets. This opens the way to tackling a variety of other problems where different sources of information may be utilized, such as in precision medicine, for instance.

The `R` package already incorporates these two possibilities.

Perspectives

Given these interesting results and possible extensions, there are other aspects that may be interesting to address in future works, with a view to improving SICOMORE further in terms of model consistency. Although the Lasso procedure is relevant for dimension reduction purposes, it may induce some biases in the multiple testing procedure used afterwards, since the variable selection step is performed before the p-values are adjusted. One way around this problem might be to use post-hoc inference for multiple comparisons [Goeman et al., 2011]. These kinds of extensions should have a positive impact on precision results.

References

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 44(2):139–177, 1982.

J. Aitchison and C. H. Ho. The multivariate poisson-log normal distribution. *Biometrika*, 76(4):643–653, 1989.

A. Alexa and J. Rahnenfuhrer. *topGO:Enrichment Analysis for Gene Ontology*, 2019. URL `https://doi.org/doi:10.18129/B9.bioc.topGO`. `R` package version 3.10.

C. Ambroise, J. Chiquet, F. Guinot, and M. Szafranski. *sicomore: selection of interaction effects in compressed multiple omics representations*, 2020. URL `http://julien.cremeriefamily.info/sicomore-pkg/`. `R` package version 0.2.1.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th Annual International Conference on Machine Learning*, pages 33–40, 2008.

H. Benjamin and T. Hothorn. *stabs: Stability Selection with Error Control*, 2017. URL `https://cran.r-project.org/package=stabs`. `R` package version 0.6-3.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 57(1):289–300, 1995.

R. L. Berendsen, C. M. J. Pieterse, and P. A. H. M. Bakker. The rhizosphere microbiome and plant health. *Trends in plant science*, 17(8):478–486, 2012.

J. Bergelson, J. Mittelstrass, and M. W. Horton. Characterizing both bacteria and fungi improves understanding of the arabidopsis root microbiome. *Scientific reports*, 9(1):1–11, 2019.

J. Bien, J. Taylor, and R. Tibshirani. A Lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe. Defining operational taxonomic units using dna barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943, 2005.

N. J. Brewin. Plant cell wall remodelling in the rhizobium–legume symbiosis. *Critical Reviews in Plant Sciences*, 23 (4):293–316, 2004.

B. J. Callahan, P. J. McMurdie, and S. P. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639, 2017.

J.-A. Chevalier, J. Salmon, and B. Thirion. Statistical inference with ensemble of clustered desparsified lasso, 2018. arXiv:1806.05829.

J. Clavel. Progress in the epidemiological understanding of gene–environment interactions in major diseases: cancer. *Comptes rendus biologies*, 330(4):306–317, 2007.

D. Clayton. *snpStats:SnpMatrix and XSnpMatrix classes and methods*, 2019. URL `https://doi.org/doi:10.18129/B9.bioc.snpStats`. `R` package version 3.10.

A. Dehman, C. Ambroise, and P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(1):148, 2015.

D. L. Donoho and Y. Tsaig. Fast solution of-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

M. Fischer, B. Strauch, and B. Y. Renard. Abundance estimation and differential testing on strain level in metagenomics data. *Bioinformatics*, 33(14):i124–i132, 2017.

G. B. Gloor, J. M. Macklaim, M. Vu, and A. D. Fernandes. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics*, 45:73–87, 2016.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

J. J. Goeman, A. Solari, et al. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

A. D. Gordon. *Classification*. Monographs on statistics and applied probability. Chapman & Hall, CRC Press, Boca Raton, Florida, United-States of America, 1999.

B. Gourion, F. Berrabah, P. Ratet, and G. Stacey. Rhizobium–legume symbioses: the crucial role of plant immunity. *Trends in plant science*, 20(3):186–194, 2015.

Q. Grimonprez. *Sélection de groupes de variables corrélées en grande dimension*. PhD thesis, Université de Lille, 2016.

Q. Grimonprez, S. Blanck, A. Celisse, G. Marot, Y. Yang, and H. Zou. *MLGL: Multi-Layer Group-Lasso*, 2020. URL `https://cran.r-project.org/package=MLGL`. `R` package version 0.6-1.

F. Guinot, M. Szafranski, C. Ambroise, and F. Samson. Learning the optimal scale for GWAS through hierarchical SNP aggregation. *BMC Bioinformatics*, 19(1):459–472, 2018.

S. Hacquard, S. Spaepen, R. Garrido-Oter, and P. Schulze-Lefert. Interplay between innate immunity and the plant microbiota. *Annual review of phytopathology*, 55:565–589, 2017.

S. S. Han and N. Chatterjee. Review of statistical methods for gene-environment interaction analysis. *Current Epidemiology Reports*, 5(1):39–45, 2018.

A. M. Hancock, B. Brachi, N. Faure, M. W. Horton, L. B. Jarymowycz, F. G. Sperone, C. Toomajian, F. Roux, and J. Bergelson. Adaptation to climate across the arabidopsis thaliana genome. *Science*, 334(6052):83–86, 2011.

M. A. Hassani, P. Durán, and S. Hacquard. Microbial interactions within the plant holobiont. *Microbiome*, 6(1):58, 2018.

J. S. Hawe, F. J. Theis, and M. Heinig. Inferring interaction networks from multi-comics data-a review. *Frontiers in genetics*, 10:535, 2019.

B. Hofner, L. Boccuto, and M. Göker. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, 16:144, 2015.

M. W. Horton, N. Bodenhausen, K. Beilsmith, D. Meng, B. D. Muegge, S. Subramanian, M. M. Vetter, B. J. Vilhjálmsson, M. Nordborg, J. I. Gordon, et al. Genome-wide association study of arabidopsis thaliana leaf microbial community. *Nature Communications*, 5(1):1–7, 2014.

S. Huang, K. Chaudhary, and L. X. Garmire. More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8, 2017.

C. M. Hutter, L. E. Mechanic, N. Chatterjee, P. Kraft, E. M. Gillanders, and N. G.-E. T. Tank. Gene-environment interactions in cancer epidemiology: a national cancer institute think tank report. *Genetic epidemiology*, 37(7): 643–657, 2013.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, et al. Pycogent: a toolkit for making sense from sequence. *Genome biology*, 8(8):R171, 2007.

D. Knights, M. S. Silverberg, R. K. Weersma, D. Gevers, G. Dijkstra, H. Huang, A. D. Tyler, S. Van Sommeren, F. Imhann, J. M. Stempak, et al. Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome medicine*, 6(12):107, 2014.

S. Lee, N. Görnitz, E. P. Xing, D. Heckerman, and C. Lippert. Ensembles of Lasso screening rules. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

Y. Li, F.-X. Wu, and A. Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2):325–340, 2016.

M. Lim and T. Hastie. Learning interactions via hierarchical group-Lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

M. Lim and T. Hastie. *glinternet: Learning Interactions via Hierarchical Group-Lasso Regularization*, 2019. URL `https://cran.r-project.org/package=glinternet`. `R` package version 1.0.10.

X. Lin, S. Lee, D. C. Christiani, and X. Lin. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681, 2013.

C. Lozupone and R. Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235, 2005.

B. Lugtenberg and F. Kamilova. Plant-growth-promoting rhizobacteria. *Annual review of microbiology*, 63:541–556, 2009.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B (Methodological)*, 72(4):417–473, 2010.

G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

K. G. Mukerji, C. Manoharachary, and B. P. Chamola. *Techniques in mycorrhizal studies*. Springer Science & Business Media, Dordrecht, Boston, London, 2002.

M. Y. Park, T. Hastie, and R. Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, 2007.

K. Pearson. Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60:489–498, 1896.

R. Pinton, Z. Varanini, and P. Nannipieri. *The rhizosphere: biochemistry and organic substances at the soil-plant interface*. CRC press, 2007.

J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

A. Rau. *Statistical methods and software for the analysis of transcriptomic data*. Habilitation à diriger des recherches, Université d'Evry Val d'Essonne, 2017.

N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower. Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6):R60, 2011.

Y. She, Z. Wang, and H. Jiang. Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association*, 113(521):445–454, 2016.

G. Srinivas, S. Möller, J. Wang, S. Künzel, D. Zillikens, J. F. Baines, and S. M. Ibrahim. Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering. *Nature communications*, 4, 2013.

V. Stanislas, C. Dalmasso, and C. Ambroise. Eigen-epistasis for detecting gene-gene interactions. *BMC Bioinformatics*, 18(1):54–67, 2017.

Z. Su, J. Marchini, and P. Donnelly. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304, 2011a.

Z. Su, J. Marchini, and P. Donnelly. *HAPGEN: version 2*, 2011b. URL `https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html`. Version v2.1.2.

S. Terrat, R. Christen, S. Dequiedt, M. Lelièvre, V. Nowak, T. Regnier, D. Bachar, P. Plassart, P. Wincker, C. Jolivet, A. Bispo, P. Lemanceau, P.-A. Maron, C. Mougel, and L. Ranjard. Molecular biomass and metataxogenomic assessment of soil microbial communities as influenced by soil dna extraction procedure. *Microbial biotechnology*, 5 (1):135–141, 2012.

S. Terrat, P. Plassart, E. Bourgeois, S. Ferreira, S. Dequiedt, N. Adele-Dit-De-Renseville, P. Lemanceau, A. Bispo, A. Chabbi, P.-A. Maron, and L. Ranjard. Meta-barcoded evaluation of the iso standard 11063 dna extraction procedure to characterize soil bacterial and fungal community diversity and composition. *Microbial biotechnology*, 8 (1):131–142, 2015.

D. Thomas. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11(4): 259–272, 2010.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 63, 2001.

Y. Tu, S. Rochfort, Z. Liu, Y. Ran, M. Griffith, P. Badenhorst, G. V. Louie, M. E. Bowman, K. F. Smith, J. P. Noel, A. Mouradov, and G. Spangenbergothers. Functional analyses of caffeic acid o-methyltransferase and cinnamoyl-coa-reductase genes from perennial ryegrass (lolium perenne). *The Plant Cell*, 22(10):3357–3373, 2010.

W. Underwood. The plant cell wall: a dynamic barrier against pathogen invasion. *Frontiers in plant science*, 3:85, 2012.

B. Wang, M. Yao, L. Lv, Z. Ling, and L. Li. The human microbiota in health and disease. *Engineering*, 3(1):71–82, 2017.

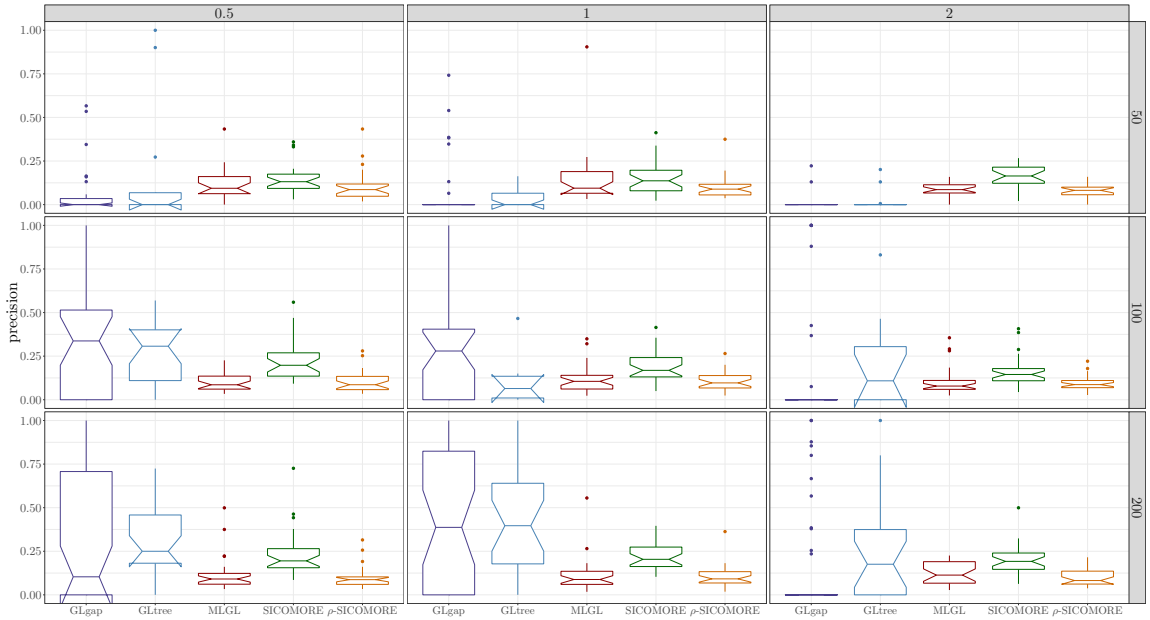
J. Wang and H. Jia. Metagenome-Wide Association Studies: fine-mining the microbiome. *Nature Reviews Microbiology*, 14(8):508–522, 2016.

J. Wang, L. B. Thingholm, J. Skiecevičienė, P. Rausch, M. Kummen, J. R. Hov, F. Degenhardt, F.-A. Heinsen, M. C. Rühlemann, S. Szymczak, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nature Genetics*, 2016.

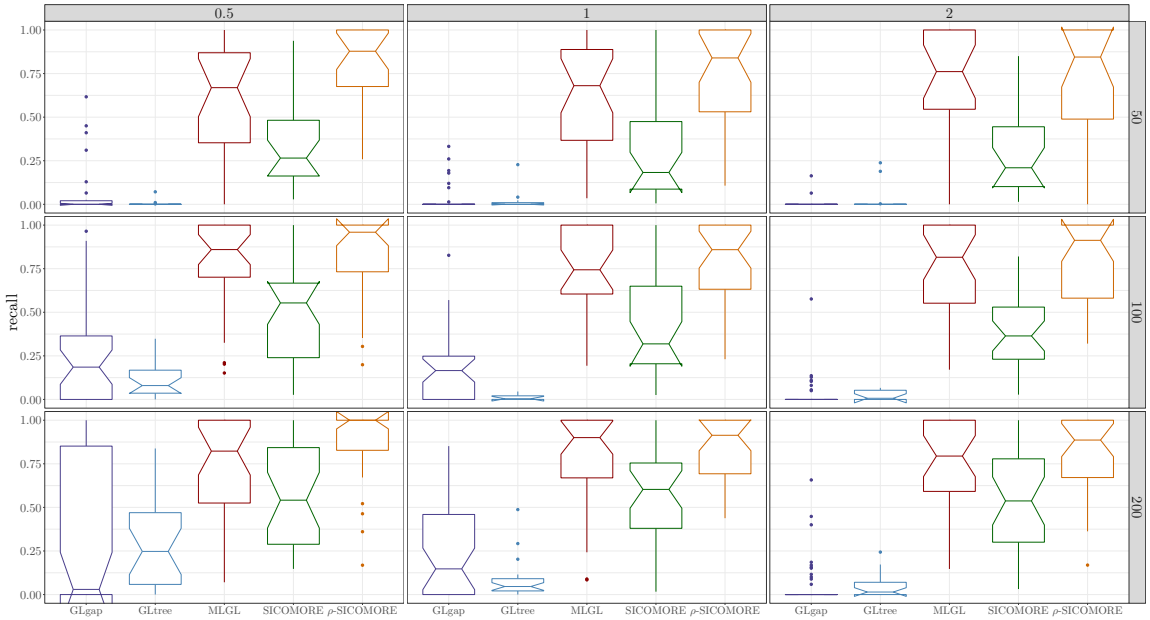
J. H. J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, 2017.

A Supplementary results

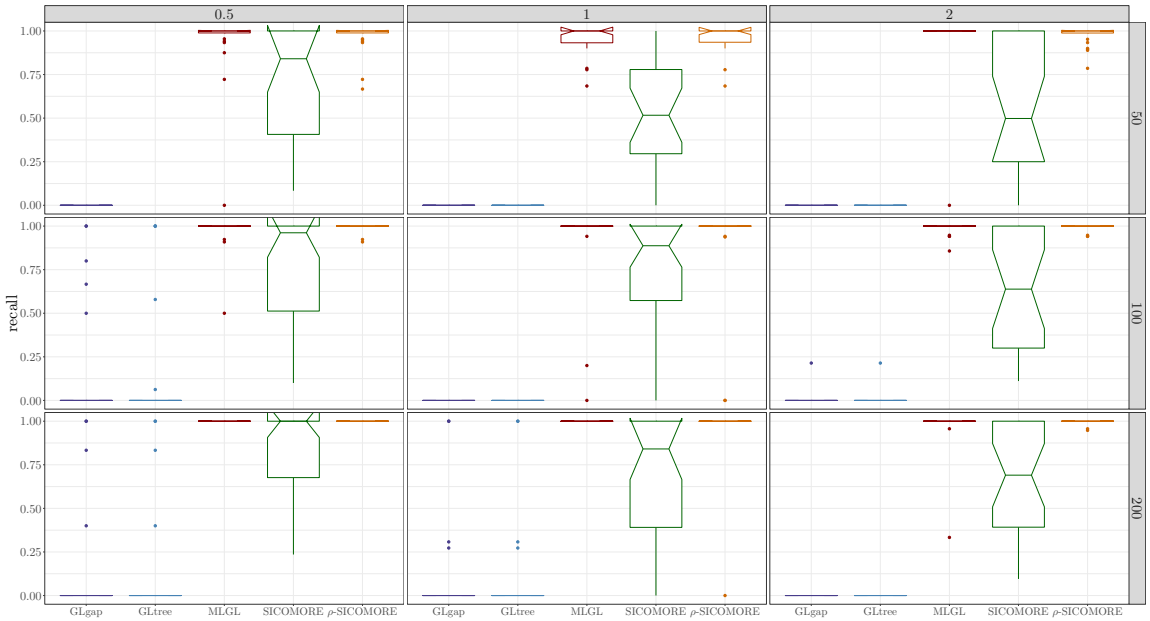


(a)

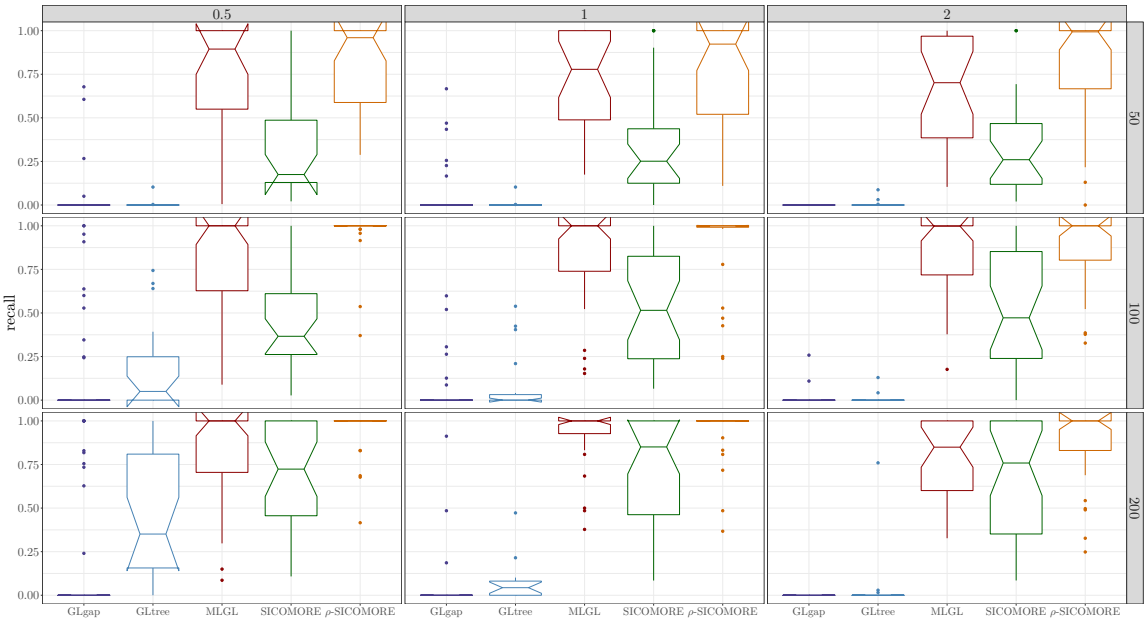


(b)

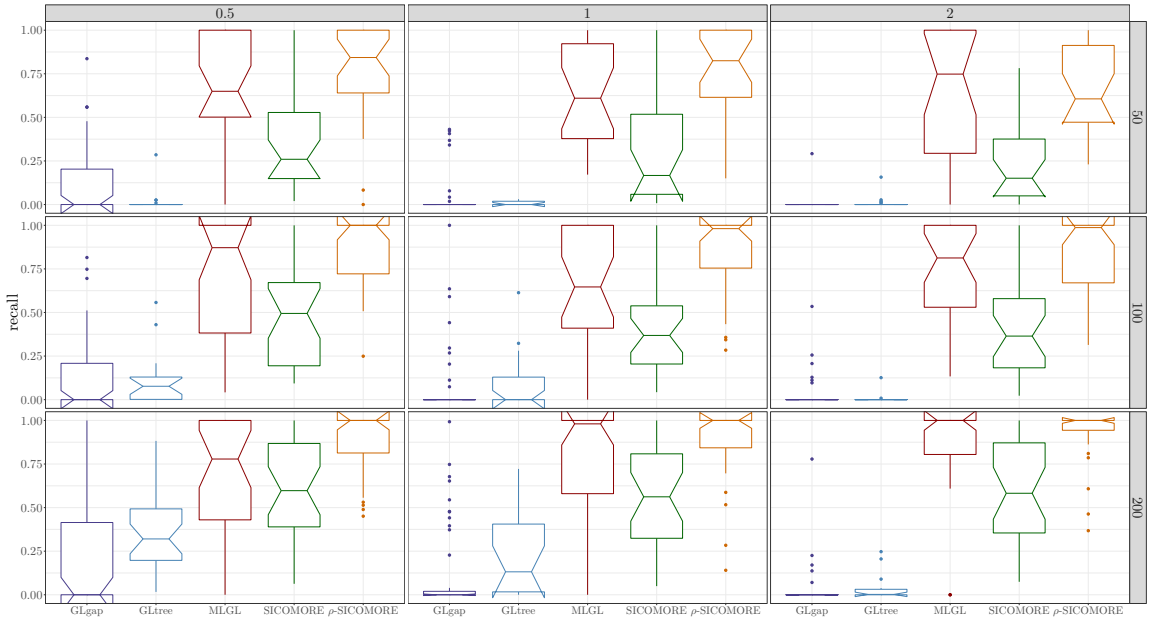
Figure 5: Boxplots of (a) Precision and (b) Recall results obtained on the numerical simulations with a Bonferroni-Holm correction for $I = 7$ blocs. The lines correspond to different numbers of observations (top: $N = 50$, middle: $N = 100$ and bottom: $N = 200$), and the columns correspond to levels of difficulty of the problem (left: $\epsilon = 0.5$, middle: $\epsilon = 1$ and right: $\epsilon = 2$). The boxplots are best seen in colors: from the left to the right, GLgap is in purple, GLtree is in blue, MLGL is in red, SICOMORE is in green, ρ-SICOMORE is in orange.


0.5
1
2
50
100
200
recall
0.00
0.25
0.50
0.75
1.00
GLgap
GLtree
MLGL
SICOMORE
ρ-SICOMORE

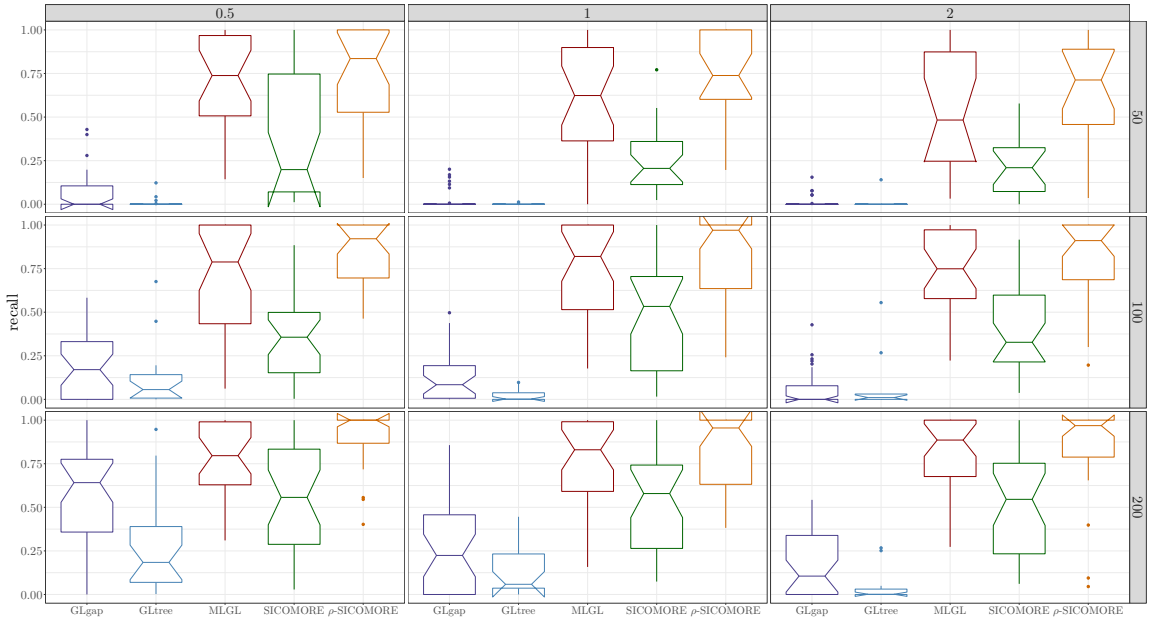
(a) $I = 1$


0.5
1
2
50
100
200
recall
0.00
0.25
0.50
0.75
1.00
GLgap
GLtree
MLGL
SICOMORE
ρ-SICOMORE

(b) $I = 3$

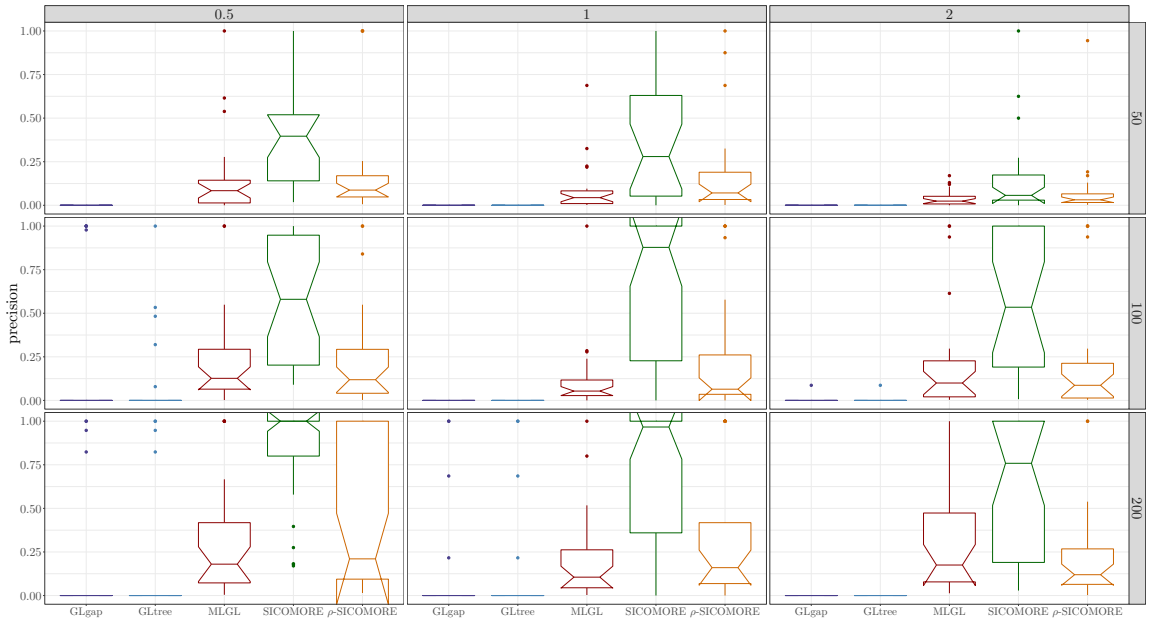


(c) $I = 5$

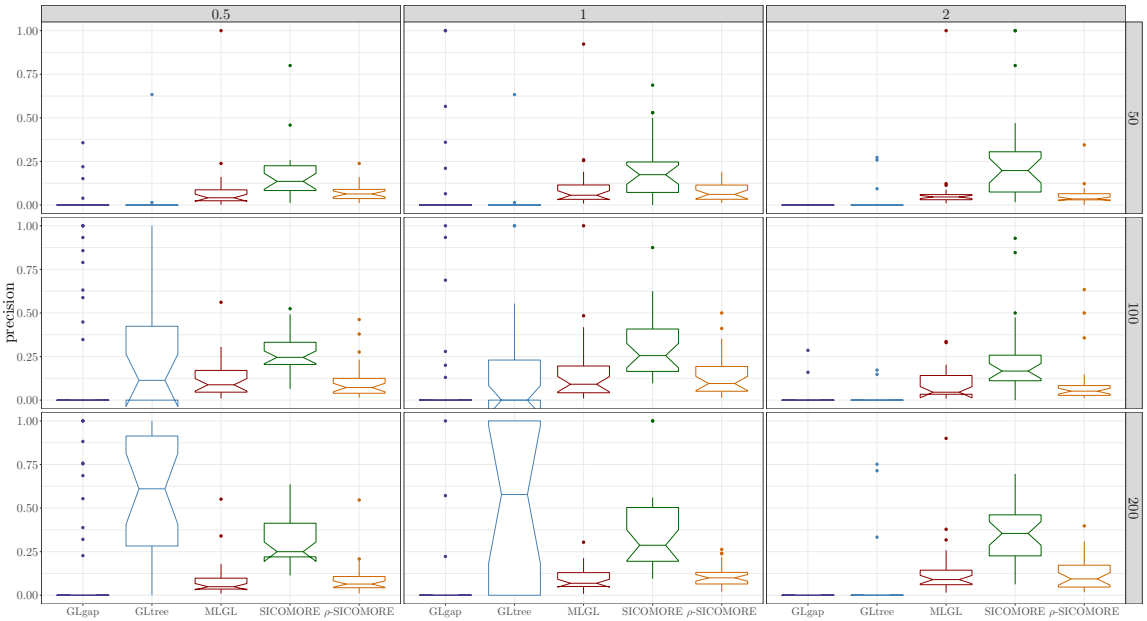


(d) $I = 10$

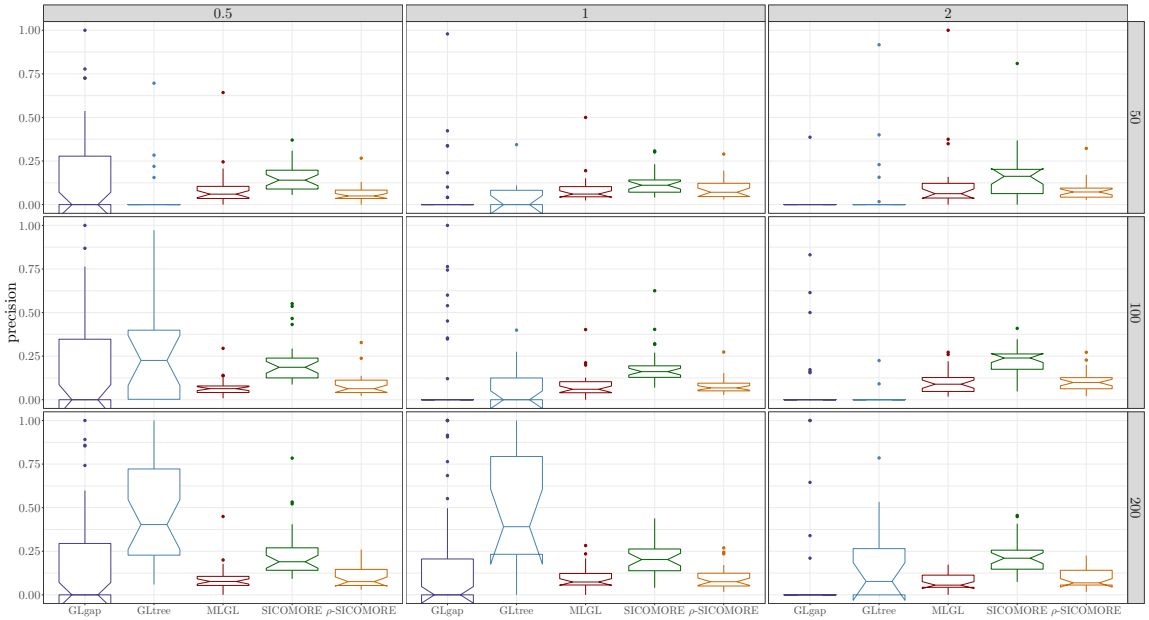
Figure 6: Boxplots for Recall obtained on the numerical simulations with a Bonferroni-Holm correction for $I = \{1, 3, 5, 10\}$ blocs. The lines show the results for different number of observations (top: $N = 50$, middle: $N = 100$ and bottom: $N = 200$) and the columns the difficulty of the problem (left: $\epsilon = 0.5$, middle: $\epsilon = 1$ and right: $\epsilon = 2$). The boxplots are best seen in colors: from the left to the right, GLgap is in purple, GLtree is in blue, MLGL is in red, SICOMORE is in green, ρ-SICOMORE is in orange.


0.5
1
2
50
100
200
precision
1.00
0.75
0.50
0.25
0.00
GLgap
GLtree
MLGL
SICOMORE
ρ-SICOMORE

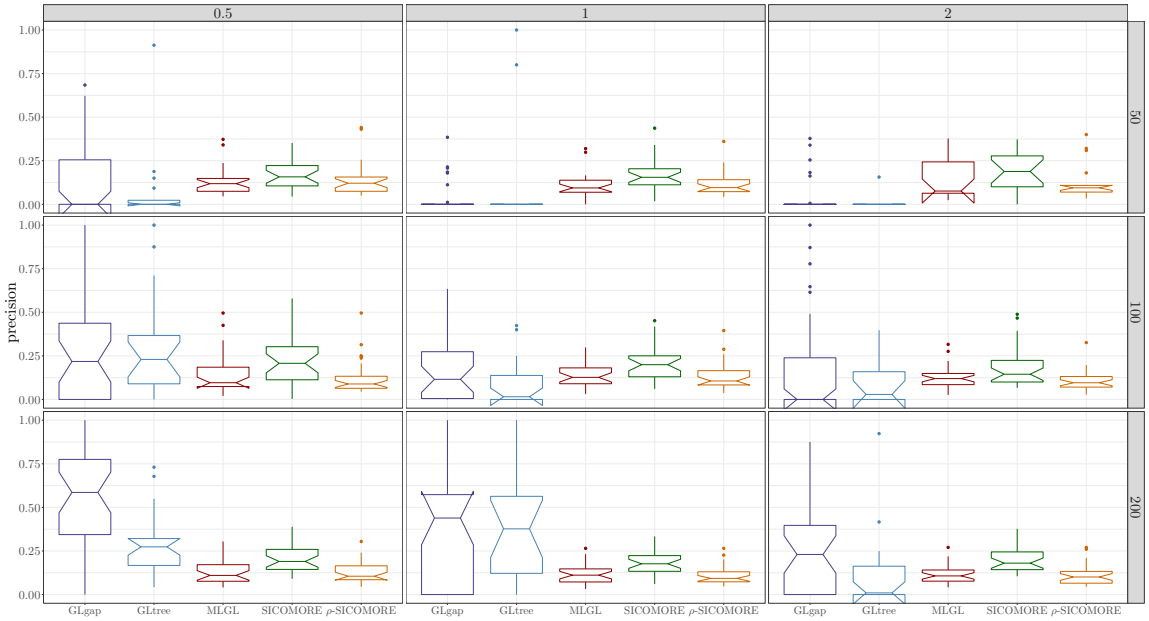
(a) $I = 1$


0.5
1
2
50
100
200
precision
1.00
0.75
0.50
0.25
0.00
GLgap
GLtree
MLGL
SICOMORE
ρ-SICOMORE

(b) $I = 3$



(c) $I = 5$



(d) $I = 10$

Figure 7: Boxplots for Precision obtained on the numerical simulations with a Bonferroni-Holm correction for $I = \{1, 3, 5, 10\}$ blocs. The lines show the results for different number of observations (top: $N = 50$, middle: $N = 100$ and bottom: $N = 200$) and the columns the difficulty of the problem (left: $\epsilon = 0.5$, middle: $\epsilon = 1$ and right: $\epsilon = 2$). The boxplots are best seen in colors: from the left to the right, GLgap is in purple, GLtree is in blue, MLGL is in red, SICOMORE is in green, ρ-SICOMORE is in orange.

B Supplementary data for the *Medicago truncatula* example

With this supplemental data, we intend to provide details to explain the metabarcoding analysis leading to the OTUs used in the *Medicago truncatula* example. This is a part of a side research paper in preparation of Anouk Zancarini, Christine Le Signor and Christophe Mougel.

Metabarcoding analysis

To assess the bacterial communities, the variable region V4 of the 16S rRNA gene was amplified using the 479F and 888R primers and sequenced using Illumina MiSeq sequencing technology (paired-end 2×250 pb). Bioinformatic analyses were done using the GnS-PIPE developed by the GenoSol platform (INRA, Dijon, France) [Terrat et al., 2012]. The details of all steps have been already described previously [Terrat et al., 2015].

After preprocessing, alignment and clustering of reads at 95% of similarity, a filtering step was carried out to check all single-singletons [5] to eliminate PCR chimeras and large sequencing errors produced by the PCR step, based on the quality of their taxonomic assignments. More precisely, each single-singleton was compared with a dedicated reference database from the Silva curated database using similarity approaches (USEARCH), with sequences longer than 500 nucleotides, and kept only if their identity was higher than the defined threshold (95%). The number of high-quality reads for each sample [6] was normalized by random selection to allow an efficient comparison of the datasets and to avoid biased community comparisons.

Then, as the analysis of microbial community richness relies on the construction of similarity clusters (called OTUs), we chose here to use OTUs to examine the distribution of 16S rRNA gene sequences in our datasets. This clustering was realized with a Perl script program that groups rare reads to abundant ones, and does not count the differences in homopolymer lengths. Finally, the global contingency table of OTUs was obtained with the samples in lines and the OTUs in columns, indicating the number of reads in each OTU for all samples. The taxonomy of each OTU was determined based on the taxonomy of all reads encompassed in the OTU. More precisely, an OTU composed of more than 90% of reads of a given taxonomy is assigned to this taxonomy. The community structure was then characterized using weighted UniFrac distance [Lozupone and Knight, 2005] calculated with the PycoGent package [Knight et al., 2007] on a phylogenetic tree computed using FastTree and the most abundant sequence to represent each OTU.

One sample was removed because of its too low-depth [Weiss et al., 2017]. The OTUs with counts lower than 41 over all the samples were filtered. The threshold of 41 was determined using the following procedure: for a threshold varying from 1 to 150, we calculated the number of OTUs whose total counts over all the samples is below this threshold. The selected threshold is the one for which the number of OTUs does not increase when the threshold increases by one. Then, the number of reads in each OTU was first summed for the three replicates for each plant genotype and a between-sample normalization was performed in order to correct for the different sequencing depth. Each sample was scaled by a size factor calculated as the ratio between the total number of counts in this sample and the mean of total counts across all samples. Finally, for each plant genotype, the number of reads were summed for OTU belonging to the same genus. OTUs that had unknown taxonomic assignment at genus level were discarded. Thus, a total of 155 samples and 329 genus were finally analysed.

All raw data sets are publicly available in the European Nucleotide Archive (ENA) of EMBL-EBI database system under project accession PRJEB25849 entitled "*Genome-wide association study of Medicago truncatula rhizosphere microbial communities and plant nutritional strategies*" with raw sequences accession (ERR2495157 to ERR2495714).

Taxonomic affiliations of OTUs

We provide a pie chart that depicts the taxonomic affiliation of the OTUs at phylum level in Figure 8. This results will be presented as a boxplot and discussed in the side paper still in preparation.

[5] Single-singletons are reads detected only once and not clustered.

[6] There are 10 000 high-quality reads for each sample.

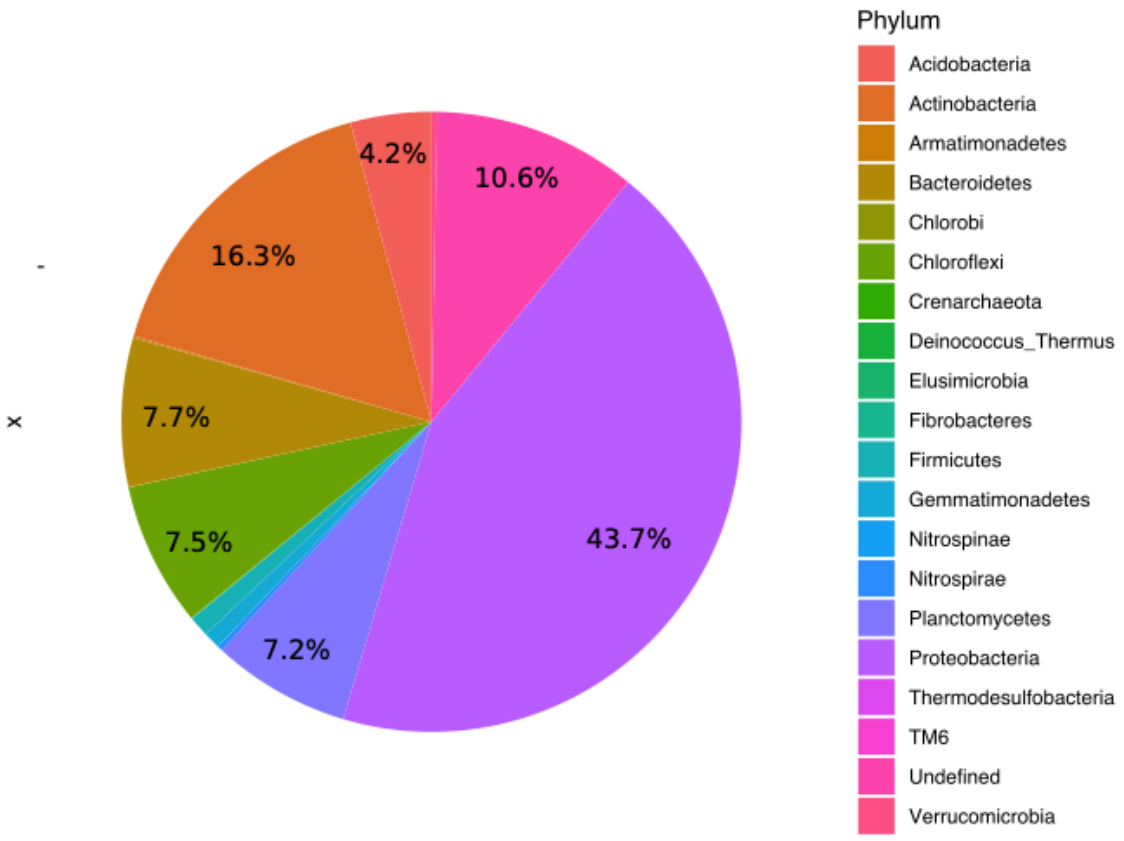


Figure 8: Taxonomic affiliation of the OTUs at Phylum level.