MASCARA: coexpression analysis in data from designed experiments

Fred T.G. White, Anna Heintz-Buschart, Lemeng Dong, Harro J. Bouwmeester, Johan A. Westerhuis, Age K. Smilde

10 January, 2024

# Abstract

Experiments in plant transcriptomics are usually designed to induce variation in a pathway of interest. Harsh experimental conditions can cause widespread transcriptional changes between groups. Discovering coexpression within a pathway of interest (here the strigolactone pathway) in this context is hampered by the dominant variance induced by the design. Minor changes in experimental conditions not controlled for may affect the plants, leading to small coordinated differences in genes within pathways of interest and related pathways between replicate plants in the same controlled experimental condition. These systematic differences between replicates within the same experimental conditions can be used to detect genes from the same or related pathway. We introduce a novel framework “MASCARA” which combines ANOVA simultaneous component analysis and partial least squares to remove the experimentally induced variance and investigate multivariate relationships in the non-designed variance. MASCARA is tested against a selection of competitors on simulated data, created to mimic a designed transcriptome study. In a coexpression analysis of a real dataset MASCARA detects several uncharacterised but relevant transcripts. Our results indicate that there is sufficient structure left in a typical dataset after correcting for experimental variance and that this residual information is useful to investigate coexpression.

# Introduction

Plant transcriptome studies typically involve the use of designed experiments which aim to induce variation in pathways of interest (POIs) by controlling one or more experimental factors. This can be done through, for example, varying the level of certain essential nutrients, knocking out/down a gene within a POI. To explore which pathways are affected by the intervention, gene expression can be measured through RNA sequencing (RNAseq). Analysis of RNAseq data generally adheres to one of two approaches: differential expression analysis (DE; ([1](#X6fa5e42e8bb074aaef1f9066b87c01d4c455181)); ([2](#ref-robinsonEdgeRBioconductorPackage2010))) or coexpression analysis (CoE; ([3](#ref-langfelderWGCNAPackageWeighted2008)); ([4](#X50bc0c3b60103ad8bb1cb40c349e8e66acca303))). DE is used to determine which genes are expressed differently between two or more experimental conditions, highlighting known or novel pathways affected by the experimental conditions. CoE, on the other hand, aims at discovering genes that are part of an already partially characterized POI. Here, the expression profiles of known genes of the POI (termed “baits”) are used to detect novel pathway members with similar expression patterns (here termed “spikes”).

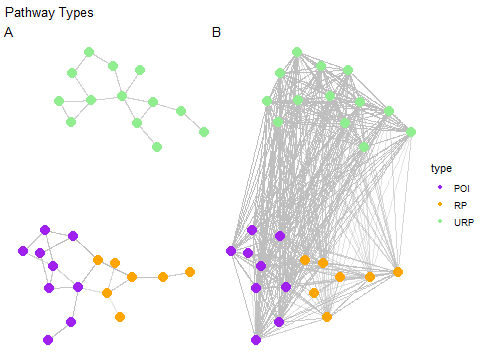


Figure 1: Pathway types exemplified, POI; pathway of interest, RP; related pathway, URP; unrelated pathway. A represents hypothetical biological pathways, B highlights the problem with detecting these pathways when dealing with data from designed experiments.

Figure 1 outlines pathway nomenclature that will be used throughout this text. Panel *A* represents hypothetical pathways including a pathway of interest (POI), related pathway (RP) and an unrelated pathway (URP). All three types of pathway are assumed to be affected by the experimental design. URPs are defined to be pathways not directly interacting with the POI. When dealing with experimentally designed data, we have to consider two key sources of variance: between-group variance and within-group variance. Between-group variance pertains to the differences between experimental groups (or conditions). This type of variance is of interest when conducting DE studies. Within-group variance, on the other hand, is the variance within experimental groups, this is not interesting for DE studies but may contain information useful to CoE. It is the excessive between-group variance that leads to the densely connected nature of panel *B* from 1. Concretising this with a biological example; in ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) a phosphate (P) starvation was applied to rice plants to induce strigolactone (SL) biosynthesis, to study the change in gene expression in comparison to control plants with normal supplies of P. The SL pathway is known to be triggered through P deficiency. However, as P is one of three key nutrients for plant growth, deficiency of this nutrient causes widespread differential expression throughout the transcriptome (([6](#ref-wangRoleStrigolactonesDeficiency2021))).

Figure 2 exemplifies between-group and within-group driven correlations from a real dataset (([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94))). The plots show the relationship between two genes of the POI (*A* and *C*) and the relationship between two genes that are not part of the same pathway (*B* and *D*). The total correlations in both *A* and *B* are high but correlation in *B* is driven by the between-group variation and not by correlations within experimental groups. The within-group correlation is lower than the total in both cases, but there is a clear trend in *C*: P- condition in red. The within group correlation (r) is higher for the first transcript pair (*C*) than the second (*D*), which makes it possible to discern between genes within a POI or RP and genes that both respond similarly to a treatment (URP). Hence, we need to distinguish *total CoE* which is calculated on the total variance and *within-group CoE* where the focus is on the shared variance within experimental groups.

From a plant biology point of view, the between-group and the within-group variance can be quite different. However, both may contain valuable biological information. The between-group variance is in principle larger because groups are the result of treatments that are usually selected because they affect the POI. Within-group variance (variance across replicates), if too large, can be an issue in DE analysis, we argue that in co-expression analysis the information present in this variance can be valuable. Variance between replicates is caused by non-controlled variations in plant development and environmental conditions. One replicate plant within a group may be slightly bigger than the others and therefore respond slightly different to a treatment. This could result in slight differences in bait genes between the replicates, but also the genes that we want to identify with a co-expression analysis. There may be many other factors that are inducing this within-group variance, but it is quite likely that if a factor induces a change in the treatment response in one replicate, that this response difference occurs throughout the entire process or pathway in that one replicate. This means that the within-group variance likely contains valuable information and by averaging out this variance, as is done in differential expression analysis, we are losing information.

Considering the within-group variance increases the strength of coexpression analysis, for example, in experiments where the intervention causes large between-group differences in many genes. Not making the distinction between types of variance results in *total CoE* associations driven by the differences between treatment groups as opposed to within group correlations. ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) used ranked Pearson correlations to determine (total) coexpression candidates. Due to such a strong effect of the experiment, many genes were affected on top of the desired SL genes. This masked the within-group correlation between the baits and the spikes (remaining unknown SL genes). This problematic issue was also highlighted recently by ([7](#ref-saccentiWhatCanGo2023)) where group sampling assumptions are investigated. Hence, both from a data analysis as well as from a biological perspective it is relevant to make a distinction between the different types of CoE. The focus of this paper will be on developing methods to investigate within-group CoE.

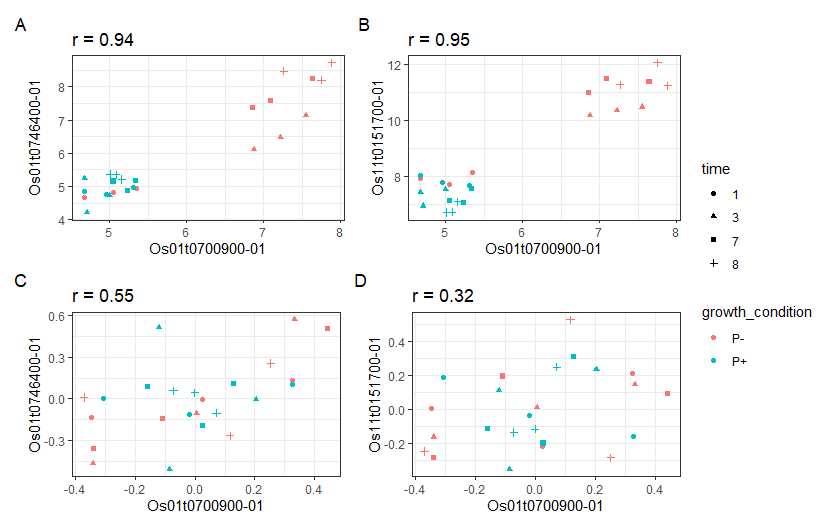


Figure 2: Correlation examples, Pearson’s correlation coefficient (r) indicated in facet titles. Data from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)). X axes, expression levels of MAX1 gene (Os01t0700900–01) in rice, Y axes expression levels of example coexpression candidates, aribitrary units. Top row A-B; *total CoE*, bottom row C-D *within-group CoE* from variance partitioning the data in A-B. Gene pairs in both A and B have strong total correlations. Correlation in A is also exhibited within a group, B has strong total correlations mainly due to the differences between groups. These examples are indistinguishable from eachother without variance partitioning.

For a successful coexpression analysis it is necessary to have a predefined set of baits that are already known to be involved in the pathway under study. Methods like ranked correlations and partial least squares regression (PLS; ([8](#Xd0e6c30a697db42abb2065af1a17da6eb418ea3)); ([9](#ref-woldCollinearityProblemLinear1984))) are able to investigate relationships between a set of baits and the other measured genes. In ranked correlations the relationships are quantified at a bivariate level. In PLS regression the relationships can be investigated in a multivariate way, in its simplest form we can see which genes are most predictive of the expression of one bait. This can be extended to a PLS2 model where relationships are investigated between a set of multiple baits and all other genes. These methods do not by default account for the different variance types discussed above. Hence, they are not directly suitable for within-group CoE.

The most widely known and applied method that investigates variance sources is the analysis of variance (ANOVA; ([10](#ref-searleLinearModels1971))). ANOVA simultaneous component analysis (ASCA; ([11](#Xaca7e11f7e5f0a0465524ca94b5740f80e57dcd))) and its extension to imbalanced data via the generalised linear model (ASCA+; ([12](#ref-thielASCAAPCAExtensions2017))) have been developed to combine principal component analysis (PCA) with ANOVA and can therefore account for experimental design in dimension reduction. ASCA is of interest due to the variance partitioning. This is typically used to estimate the effect matrices that are calculated from the design and particularly the within-group variance structure which is maintained in the residuals of the model. Both PLS and ASCA have been used for DE (([13](#ref-augerGeneCoexpressionNetwork2022)), ([14](#ref-renRNABindingProteins2023)), ([15](#ref-nuedaDiscoveringGeneExpression2007)), ([16](#ref-jarmundALASCAPackageLongitudinal2022))) and CoE (([17](#ref-guanSRGSSparsePartial2022)), ([18](#ref-nuedaFunctionalAssessmentTime2009))). ASCA and PLS have been combined before in ([19](#ref-thissenImprovingAnalysisDesigned2009)) creating ANOVA-PLS, however the goal of their study was not coexpression analysis. Neither of these methods are optimal by themselves for CoE in the context of large between group variance with known baits.

The baits are assumed to be part of the same POI and the remaining uncharacterised spikes (POI or RP genes) are assumed to be numerous. Therefore, both the baits and the remainder of the data could be reduced to linear combinations in order to investigate relationships between a set of baits and putative spikes. We introduce a new data analysis method called multivariate ASCA residual analysis (MASCARA). MASCARA is motivated by four key factors; (1) the presence of dominant between-group variance caused by the setup of the experiment, as well as, (2) ambient variance (structured effects caused by undocumented environmental factors i.e. temperature, light intensity or soil water content) within the replicates of each experimental group, (3) the need to investigate multivariate relationships and (4) the low number of samples. The method capitalises on ASCA to account for an experimental design and the multivariate analysis capability of PLS2 to find uncharacterised POI genes using a set of baits. Instead of focusing on the variance induced by the experimental design, as is a typical use of ASCA, this variance is removed, and the remaining residual variance is further analysed to investigate the multivariate relationships between baits and the putative spikes in terms of within-group CoE.

To explore the utility of MASCARA, the method will be compared to ASCA, PLS2 and correlation analysis illustrated using an example RNAseq dataset (from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94))). Furthermore, this dataset is used as a guide to several simulation studies to validate the development of a novel approach to coexpression analysis. For these simulations we emulate both different levels of between- and within-group variance. We test the effects of size of between group differences, number of URP differential genes, measurement noise (random unstructured technical variance), structured ambient variance and number of replicates. The following sections seek to investigate which statistical methods are best suited for POI elucidation when a set of baits is already known and illustrate the benefit of MASCARA. Especially in data coming from experiments designed for differential expression analysis.

# Materials and Methods

## RNAseq - Rice strigolactone dataset ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94))

In order to guide the simulations we illustrate the desired characteristics with a real dataset. The real data are a subset of samples from a study in which the root transcriptome of rice plants is measured under two growth condition (P+/P-) and measured at four time points (1, 3, 7 and 8 days post treatment). At each combination of growth condition and time point, the root transcriptome of three new plants was harvested for RNA sequencing. In this dataset, the curated list of SL pathway genes (contained in Supplementary Table 1) shows a specific profile: low variance baseline expression across all time points in the P+ condition but activation and increasing expression over time with P starvation, data were generated and preprocessed to variance stabilised counts as per ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)).

Figure 3 shows the gene expression levels of 9 strigolactone pathway genes along with 9 highly differential genes that are induced by P deficiency, but are not part of the SL pathway. While the expression is low in P+ condition a clear increasing time profile can be observed in the P- condition, indicating an interaction between the growth condition and the time factors.

What is notable here is the replicate specific patterns that are apparent amongst the SL genes and largely absent between the SLs (purple) and the non-SL differential genes (green). For example, at time point 3, most SL genes are expressed lowest in replicate 1 and highest in replicate 3, while at time point 7 replicate 2 has a higher SL gene expression. These within-group patterns are either not shared at all between the SL genes and the other DE genes or to a much lesser extent. This is due to the other genes not directly being part of the same pathway. There are of course other pathways and thus there are still some subtle replicate patterns in the non-SL genes but this effect is much clearer within our pathway of interest. Figure 4 shows the output of the ASCA decomposition in which the PCA scores and loadings of the combined effect of the Growth condition factor and the interaction of Growth condition and Time is explored. The score plot shows that the replicates of each condition are clustered while the difference between the conditions is much larger, as expected. While the 9 SL transcripts have higher loadings on the 1st PC, there are many non-SL transcripts with similarly (or more) extreme loadings. Thus many differentially expressed genes are found important that are not part of the SL pathway, in other words ASCA does not detect SL (POI) genes from others that are similarly or more differentially expressed but not part of the SL pathway (URPs).

This marks the goal of this paper, which is to detect within-group coexpressed genes (POI/RP) with a set of baits preferentially to other differentially expressed genes that are not part of the same pathway (URPs, with a different within-group structure).

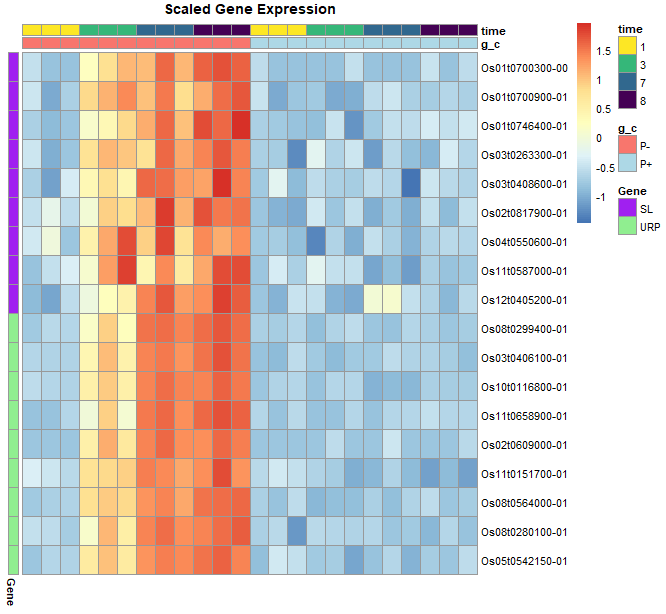


Figure 3: Real data overview. RNAseq of rice root. Heatmap of SL pathway and example highly differential genes. Data and preprocessing from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)), genes autoscaled i.e. blue indicates no/low expression and red indicates higher expression. IDs with functional annotations in supplementary table 1.

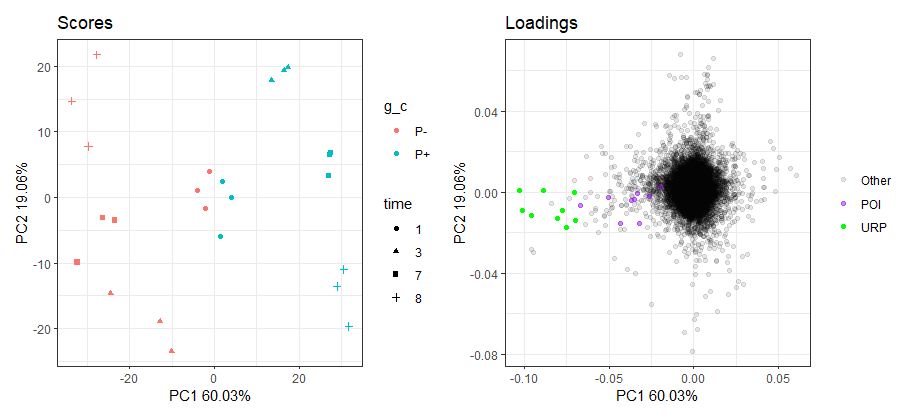


Figure 4: Scores and loadings of the ASCA model of the combined effect (condition + condition:time interaction), model indicates samples and genes relative positions in the designed variance. *g\_c* indicates growth condition, *time* indicates days after starvation induction.

## Simulations

The problem in the real dataset is that the nutrient stress applied causes a major effect on many other genes besides our baits and spikes. This widespread transcriptional regulation can mask the coexpression patterns of interest due to the large number of genes that are upregulated under the stress condition, making it more difficult to detect genes that are responding to a lesser extent, for example the SL pathway.

The simulations are built upon the results of the ASCA analysis of the combined effect of growth condition and the interaction of growth condition (g\_c) and time. As can be seen in Figure 4 there is clear separation between samples due to the experimental factors growth condition (g\_c) and time. The variation between the replicates is only small. The green loadings of URP genes are strongly related to the experimental variation due to g\_c. The simulated data is created by simulating experimental variation consisting of the main time effect and the combined effect (), and the Ambient/Residual variation which exists of the structured ambient variance () and residual variation () representing measurement error, modeled with normally distributed random values .

Each of the sources of variation except the residual variation is simulated as low dimensional variation created from 2 scores () and loadings (). For the experimental variation, the scores matrices and each consist of two orthogonal score vectors and their relative levels are indicated by the colour. The different colours in Figure 5 represent that the scores are orthogonal to each other. Each small block consists of the same score value for each of the replicates within that condition. For the Ambient variation, the scores are different for each replicate plant. The values in are obtained from a normal distribution with mean 0 and variance 1, i.e. . The matrices represent the simulated scores for their respective factors. For blues, reds, greys and golds we have 4 different shades representing 4 unique values per factor level (total 8 unique values where reds and golds are sign flipped blues and greys respectively). Similarly oranges and pinks represent a growth condition independent time profile with four unique values.

In the loading matrices , the different coloured blocks indicate different types of genes. The green block represents the bait genes of POI. The black genes represent “unknown” genes from POI or RP. Yellow and purple represent genes from URP that are affected by the experimental variation in different ways. The grey blocks represent very small background variation for those genes, while the white blocks represent that those genes are not affected by that source of background variation.

In the baits, spikes and other DE genes (all genes from POI, RP and URP) along with the remaining first half have similar positive values (represented by light grey) where the other 1000 have positive values in (grey), these are created to emulate the reality of the real data in which many genes are slightly correlated with eachother through the course of development, as well as to account for some nonlinearities which are otherwise ignored by our model. The remaining elements of and are populated with zeroes. only has values for the baits (green) and spike (black) genes. The loading values for the baits and spikes are obtained as , while the loadings for the other genes are 0.

To finalise, the response values all have an intercept . Similarly to the real data, the gene expression values for each experimental condition are simulated with three independent replicates per time-condition-combination. Each simulated data Y set thus consists of 24 samples and 2000 genes.

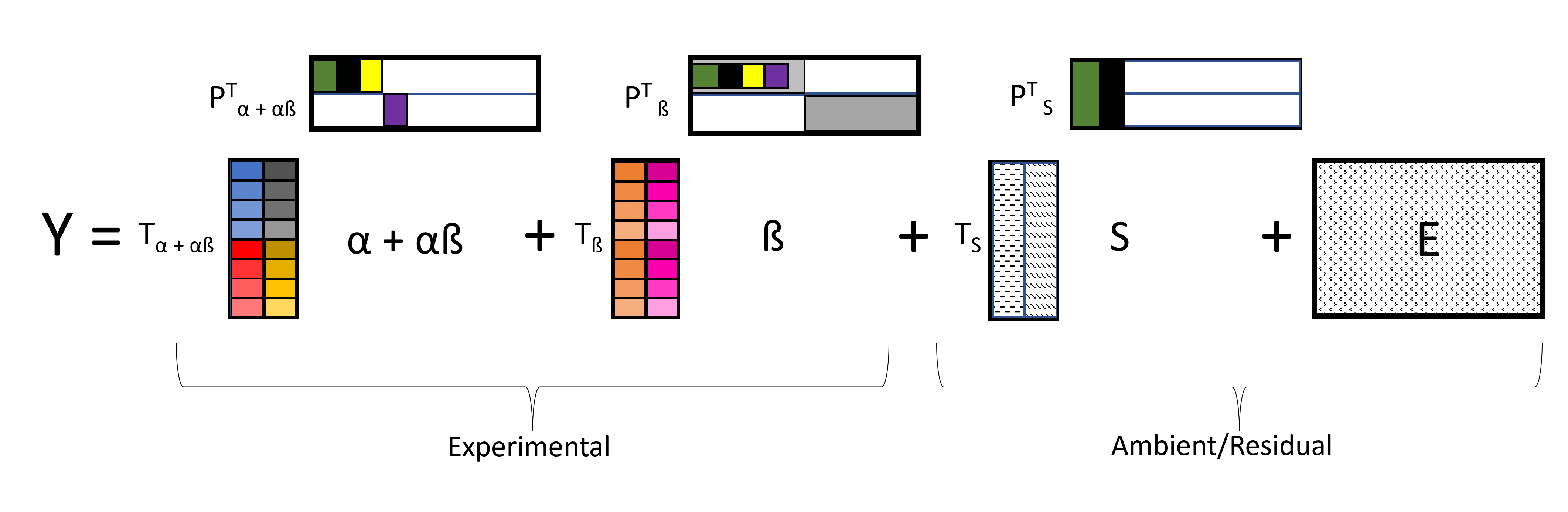


Figure 5: Coexpression simulation, simplified overview. Colours correspond to the same genes in the different loading vectors (*P*). Black; baits (known POI genes), green; spikes (unknown POI genes), yellow; DE URP genes, purple DE URP genes with a different profile. In the score vectors (T); Blues, reds, greys and golds represent experimental conditions, oranges and pinks; time. Each of the textures in the ambient/residual part represent a distinct normally distributed set of random values.

## Data simulation 1: Combined effect size and number of differential genes

The first simulation aims to determine how well the methods perform to find coexpressed genes in the presence of a dominant between-group variance structure and increasing numbers of differential URP genes (represented by the yellow block in Figure 5) that only have high between-group variance but no within-group variance.

In this simulation we have a set of 4 baits, 12 spike genes and a variable number URP genes. The issue to address is the presence of many differential URPs; those that have strong between group correlations but no within group correlation. We incrementally increase the number of genes in the yellow block and decrease number of white genes with 0 loadings. As well as the number of DE URPs we also independently control the size of the experimental variance through the parameter *d* in Equation in (1). This approach controls the size of the experimental variance as well as the number of genes responding while keeping a fixed amount of ambient or within group variation.

## Data simulation 2: Replication

The second simulation focuses on the effect of a different number of replicates. We investigate a range of values for sizes of between experimental variance (*d*) and random noise variation (*l*) across random structures (which correspond to 20 repeated simulations of randomised values for all matrices).

On these 20 datasets, per condition, the replicate number was also varied: the number of replicates per experimental condition were sampled from a maximum of 15.

With this setup we test the feasibility range of MASCARA with respect to the ratio of 3 key variance parameters namely: 1. Within group correlation structure 2. Between group correlation structure 3. Random noise

Here we set a fixed size of within group correlation structure () as well as number of non-pathway DE genes to 30 (yellow bar size Figure 5) across all iterations of this test. We vary the sizes of the other two variance parameters (*d* and *l*). Furthermore we also show the effect of replicate number and demonstrate that MASCARA (along with the other methods) can be applied to data with generic profiles, not limited to the specific structures that were used to demonstrate the differences in coexpression analysis with simulation 1. Here we create 20 datasets with random combined effect and within group correlation structures.

## Selection of genes of interest

### Ranked correlations

Pearson’s correlation coefficient is calculated between all bait genes (b = 1..B) and all other genes (). Ranks are assigned for each such that the highest correlation for each bait gene b receives the lowest rank. These ranks are averaged across all baits B (see supplementary equation (14)).

### ASCA+

ANOVA simultaneous component analysis (([11](#Xaca7e11f7e5f0a0465524ca94b5740f80e57dcd))) involves splitting data into effect matrices which are defined by the experimental design structure, after which a PCA is applied to the effect matrices or combination thereof. In ASCA+ this variance partitioning is achieved through the generalised linear model as described in ([12](#ref-thielASCAAPCAExtensions2017)). ASCA+ addresses the issue of biased effects matrices in unbalanced designs. The following is a summary of the algorithm description provided in ([12](#ref-thielASCAAPCAExtensions2017)).

The multivariate generalised linear model (GLM) can be expressed as:

where is the response matrix of dimension , is the model matrix of dimension , containing the sumcoded indicator levels for each of the levels for each factor and their interactions, is the parameter matrix of dimension and is the error matrix of dimension .

Through Ordinary Least Squares (OLS) we can obtain the (unbiased) estimators of the model parameters:

The effect matrices for different terms of the model can be obtained as:

where: is the effect matrix corresponding to effect , is the matrix of estimated parameters obtained from OLS and is a new model matrix obtained by keeping in only the block and replacing all other columns with zeros as per:

The above is achieved through the sum coded indicator matrix as illustrated further in ([12](#ref-thielASCAAPCAExtensions2017)).

Furthermore, the matrix is estimated as the residual matrix of the model:

Throughout this text when referring to the calculation of ASCA/ASCA+ model on simulated or real data the following model is calculated:

With genes selected from:

### PLS regression

Partial Least Squares (PLS2) is a multivariate regression technique used for modeling relationships between two sets of variables. Here we split our data matrix into sub-matrices;

The response matrix contains only the baits (relating to the green bar in the loadings of Figure 5) and the predictor matrix contains the rest of the data, this can then be expressed as:

This model is estimated in our work using the SIMPLS algorithm (([20](#ref-dejongSIMPLSAlternativeApproach1993))) as follows:

1. Initialize to and to .
2. Iteratively perform the following steps for :
   1. Compute the singular value decomposition (SVD) of to obtain the left and right singular vectors; weights and respectively.
   2. Project and onto the first left and right singular vectors (, ) to obtain the scores and .

Under the conditions of orthogonal scores; and normalised weights; and .

PLS2 aims to maximize the covariance between the scores and at each iteration. In the following simulations and application the number of components calculated by PLS2 is set to 2 to create an interpretable model.

### Variable importance in projection (VIP)

In this work genes are selected in both ASCA+ and PLS through VIP which is a way to measure the contributions of each variable to the underlying models. According to:

Where k=1…K represents the number of components in the respective models, j indexes a gene, is substituted for the loadings or weights (for ASCA+ and PLS respectively) and VE is the variance explained, where is the amount of variance explained by component k and is the variance explained by the whole model. In ASCA+ we calculate the VIP scores in the latent space of as described in ([12](#ref-thielASCAAPCAExtensions2017)). In PLS we use the VE from the latent space of using the weights () from the latent space of .

### Multivariate ASCA residual analysis (MASCARA)

MASCARA combines the effect estimation with GLM (from ASCA+) with PLS2. The method entails estimating effect matrices with the GLM step of ASCA+ as described above to obtain the estimated residual matrix . The variance explained by all effect matrices of the experimental factors and interactions is removed. What remains in this matrix is individual variance of each plant that is assumed to contain not just random technical variance but some form of structured variance induced by ambient (uncontrolled) effects. The matrix from equation (7) is substituted as into PLS2 after being split into baits and remainder as per equation (10).

The genes are ranked based on target projection (([21](#Xf6b94e0f73326edb35ae19b864793f61a792cea))) of the loadings onto the mean vector of the bait loadings in 2 components: where is the mean of the component of the PLS2 model. The target projected loadings are calculated such that . Higher values in the target projection equate to stronger positive associations with the center of the baits. Large negative values can also be interesting candidates for strong negative associations, although non positive associations are not considered here. This approach allows a general direction in the PLS space to be defined by multiple baits and assumes that these baits are highly correlated with one another.

## Performance evaluation; Log2 geometric mean rank

To evaluate performance we use log2 geometric mean rank (log2(GMR); ), defined as:

for spikes. A lower indicates better performance.

# Results

## Simulation 1: Combined effect size and number of differential genes

For this simulation we have some experimental variance for the combined effect and time ( and respectively) a fixed amount of random noise and a fixed amount of structured ambient/residual variance , within which our 4 baits and 12 spikes share a correlation structure. This reflects the situation expected from the real data example - drastic effect caused by nutrient deficiency, which also activates the SL pathway. The POI/RP genes are expected to share some variance that is independent to the experiment - the 4 baits and 12 spikes load on to the structured part of the noise .

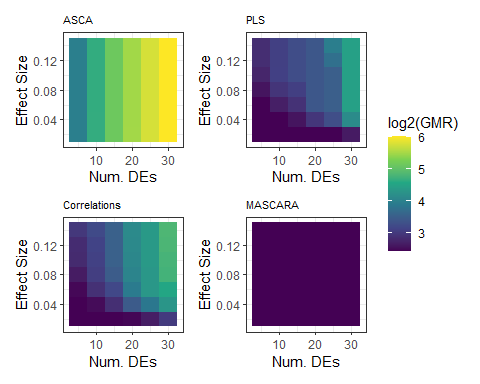


Figure 6: Results of simulation 1 for all 4 methods. Log2 geometric mean rank (log2(GMR)) as a function of the number of differentially expressed URP genes (x-axis) and combined effect size (y axis); (), lower log2(GMR); dark blue, indicates better performance. ASCA model calculated per equation (8), PLS model per equation (11).

We can see in Figure 6 that MASCARA was neither affected by the presence of an increasing number of non-pathway DE genes nor the size of this effect. Correlations, ASCA and PLS, however, all showed the expected breakdown in performance due to the injection of more DE URP genes. This was the expected result as the experimental between group variance is removed with MASCARA.

## Simulation 2: Replication

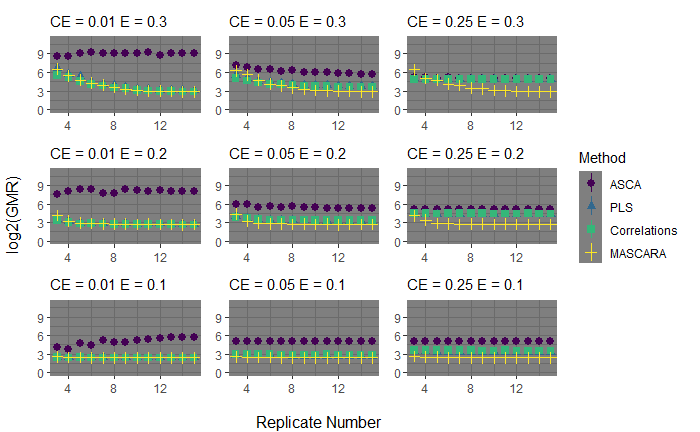


Figure 7: Simulation 2 results. Replications and variance type feasibility ranges. X axes; replicate number, y axes log2 transformed geometric mean rank. Each facet is created with set parameters for combined effect size () and random gaussian noise size (). Dots are median performance across 20 random structures. ASCA model calculated per equation (8), PLS model per equation (11).

It was expected that all geometric mean ranks converge towards a minimum with increasing numbers of replicates, see 7. Perfect detection of 12 spikes in top 12 candidates is achieved with a log2 geometric mean rank of 2.4, MASCARA was able to achieve this, however, the other methods were perturbed by the presence of the 30 differential URP genes. With more replicates MASCARA always outperforms competitors. With higher values of *d* (larger combined effect size) (CE) MASCARA outperforms the other methods but its performance is affected by amount of random noise (E) size indicated by *l*.

In simulation 2 we fixed the number of differential URP genes to 30, however this is less than 2% of the total number of genes in each dataset. In the real data example more than 10% of of the total number of genes are differentially expressed. As shown in simulation 1 (Figure 6) higher numbers of DE URP genes results in higher log2 geometric mean rank (log2(GMR)) of the spikes for correlations, ASCA and PLS. MASCARA is designed specifically to mitigate the effect of large numbers of DE URP genes. Here we show that this is achieved.

# Real Data Applications

It is apparent that MASCARA has a particular use case for coexpression analysis in data where there is a lot of experimentally induced between-group variance across many genes as well as some within-group correlation structure shared between POI genes that is caused by independent non-controlled ambient factors.

Here we illustrate an application of MASCARA to the dataset from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)). For this we take 4 core strigolactone genes (Os11t0587000, Os04t0550600, Os01t0746400 and Os01t0700900 of the 9 used for illustrations in Figures 3 and 4) as baits and investigate their relationship with top candidate genes.

In the real data from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) some plants were subjected to massive stress with the phosphate limitation. This is, therefore, reflected by large differences in the transcriptome compared to those with a normal supply of P. This limitation affected many more genes than those directly involved in SL biosynthesis. Furthermore, there appears to be a stronger relationship amongst the SL genes in the ambient variation compared to their correlations with other DE URP genes, as displayed in Figure 8 where the blue distribution indicates correlations between residuals of the 9 POI transcripts from Figures 3 and 4 and the red indicates correlations between residuals of these 9 and top DE URP genes.

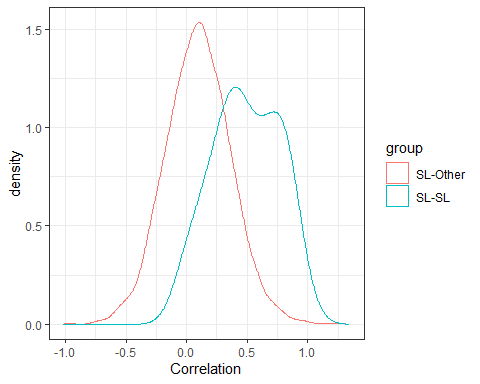


Figure 8: Distributions of Fisher transformed Pearson correlations (supplementary equation (15)) in residuals from the ASCA model on the data from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)). Blue indicates correlations between SL pathway genes, red indicates correlations between SL pathway genes and top 10% DE URP genes (upregulated in P- over P+).

Some example top coexpression candidate genes are shown in Figure 9. The problem of dominant between group variation is again exemplified here. Several of the top candidates detected by correlations (ASCA and PLS, results not shown) show no indication of within group correlations. Candidates selected with MASCARA, on the other hand, consistently show within group correlations with the baits. Furthermore, the strigolactone pathway, which is still under active investigation, has several uncharacterised biosynthetic steps. These are thought to be carried out by cytochrome p450s. The top candidates from MASCARA contain at least three (putative) cytochrome p450 genes that have not yet been biologically investigated in SL biosynthesis, as well as, a methyltransferase and several putative carotenoid associated genes. MASCARA also detected genes linked to Gibberellins (GA) preferentially to general phosphate starvation genes. The GA pathway has been linked to the strigolactones (([22](#Xbd25994ee0f44976988919f4c291e7808fe56a3)), ([23](#Xea043edabf8aaad9d19b1e892525d9fffe3a696))). The interaction between these two pathways has been characterised recently under nitrogen limitation by ([24](#X2e16474dbff84e9966c63a78f86e923d201544c)).

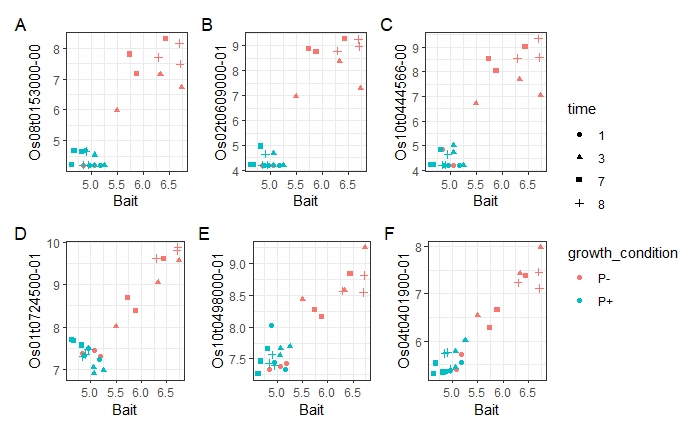


Figure 9: Selected candidates from A-C ranked correlations and D-F MASCARA. X axes; transformed expression level of one of the bait genes (CCD8), y axes; expression levels of three example coexpression candidates for each method (see ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) for preprocessing details).

# Discussion and conclusions

In this article we introduced multivariate ASCA residual analysis (MASCARA) as a method for coexpression analysis in data containing a dominant variance structure from a known (experimental) source. MASCARA estimates effects through the GLM with a design matrix to remove the variance between conditions, then the residual matrix can be decomposed through PLS2, with feature ranking through target projections. This framework enables the investigation of multivariate relationships between a set of bait genes and all other genes in the dataset. MASCARA thrives on the fact that within the typical experimental setup small systematic variance exists between replicates due to variations in ambient factors that affect the plant, and therefore the variance structures in the resulting data.

Through simulation studies we showed the conditions under which MASCARA is suitable for coexpression analysis and compared it to a selection of other coexpression methods (ASCA+, PLS2 and correlations). These methods were applied to data from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) exploring the strigolactone pathway and simulated data based on the structure of the Haider data. In a coexpression context, within certain assumed parameters of variance structure (dominant between-group, existing within-group), MASCARA outperformed competitors. Ranked correlations, ASCA (([11](#Xaca7e11f7e5f0a0465524ca94b5740f80e57dcd)), ([12](#ref-thielASCAAPCAExtensions2017))) and PLS (([9](#ref-woldCollinearityProblemLinear1984))) are all affected by dominant experimental between-group variance, which is a common occurrence in experimental data, exemplified here by the dataset from ([5](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) but discussed elsewhere (([25](#Xcc71cd2df4032d63494e28c6a7921129da1c090)), ([26](#ref-zhaoWeightedGeneCoexpression2010)), ([27](#X3e57ed0fe141b4b17e403de16a66af40da9b626))).

The feature selection stage of MASCARA uses an average vector constructed from the loadings of the baits onto which the loadings of other genes are projected. This approach is similar to target projection in ([21](#Xf6b94e0f73326edb35ae19b864793f61a792cea)) although we construct the target vector from averaging the loadings in a PLS2 model as opposed to already having it defined with the univariate response in PLS1. There are many established methods for feature selection in PLS which include Variable Importance in Projection (VIP) and Selectivity Ratio (([28](#ref-rajalahtiBiomarkerDiscoveryMass2009))), however these methods focus on the importance of variables in the construction of the model rather than a directional association with the response block as with the target projection stage. With this averaged target projection approach, as an extension of Kvalheim’s target projection in PLS1 models ([21](#Xf6b94e0f73326edb35ae19b864793f61a792cea)), we can use the information in a set of baits to define a direction in the latent space which can then be used as an axis to determine the relatedness of other genes. This assumes moderate to high correlation in the residuals of the baits, which is a realistic assumption in coexpression analysis although should be verified in the selection of baits.

Our analysis framework does not consider different correlation structures within the residual variance, this could be an issue especially in the cases similar to ([29](#Xe7babcb509b3ef554dc77224d7bc0c8531e51b3)) where different combinations of experimental factors show different within-group correlation structures. This type of relationship can be seen in Figure 9 where the P- experiments show high positive correlation between the genes of the same pathway while the P+ experiments show a negative correlation. With MASCARA these will not be preferentially detected, as the focus is on detecting positively associating features. Ideally there would be enough samples to calculate robust correlations per group, to account for possible rewiring under different conditions, but in our case we focus on experiments where there are generally not enough replicates to do this. Experiments with higher replication number are always desirable from the point of view of data analysis. However, this is not always feasible due to time and space constraints. We tested the effects of replication number and found that MASCARA can outperform its competitors with lower replication, especially in data containing many differential genes between experimental conditions. This advantage holds as long as there is a realistic amount of structured variance caused by undocumented ambient factors.

In coexpression analysis, there is no need for class balance, it is in fact better to have more data points in the class where the pathway is induced (P- in the example dataset). When the POI is only active in the case condition then the control samples are mostly redundant. With more samples in the condition of interest the reliability of the coexpression analysis will improve. The MASCARA approach indicates that the typical experimental design used for DE is not optimal for coexpression studies. The fact that the residuals need to be isolated to find within group variance confirms the notion that coexpression analysis warrants different experiments than those used for DE studies. If quantification of differential expression is also needed, for example to confirm the pathway under study is affected by the treatment, but the main focus of research is coexpression, then it is an option to have just a few samples in a control condition to confirm the experimental treatment has caused the desired effect. This was not tested explicitly here, however if pathway genes are not active in a given condition then this set of samples does not provide any useful coexpression information.

This work has outlined and tested a basic use of MASCARA where the relationship between baits and unknown pathway genes is predefined to be positive. The method is not limited to the calculation of positive associations, however performance in detecting negative correlations (indicative of inhibitory or negative feedback processes) or more complex relationships has not been tested within the current simulations. Our approach was based on the structure of the study by Haider with RNAseq measurements of rice plants, however the method is applicable across other organisms and omics types. This holds under the assumption that the set of baits have a multivariate relationship between themselves and multiple other (unknown) genes. These genes are also related to the baits, as in our partially characterised strigolactone pathway example. In untargeted metabolomics datasets under similar experimental conditions these assumptions are also likely to be met.

Taken together our results indicate that the lack of ability to discern different types of variance in a dataset can compromise the detection of true coexpression patterns. MASCARA accounts for the different types of (experimental and ambient) variance and enables finer scale detection of within group variance structure.

# Acknowledgements

We acknowledge funding by the Dutch Research Council (NWO/OCW) for the MiCRop Consortium program, Harnessing the second genome of plants (Grant number 024.004.014; to HB, LD, AKS, JAW), the Dutch Research Council (NWO-TTW grant 16873 Holland Innovative Potato; to HB and LD) and the Data Science Centre of the University of Amsterdam (to FW). FW would like to thank Frans van der Kloet, Fentaw Abegaz, Roel van der Ploeg for statistical discussions and Imran Haider for consultation on the strigolactone pathway.

# References

1. Anders S, Huber W. [Differential expression analysis for sequence count data](https://doi.org/10.1038/npre.2010.4282.1). Nat Prec. 2010 Mar 15;1–1.

2. Robinson MD, McCarthy DJ, Smyth GK. [edgeR: a Bioconductor package for differential expression analysis of digital gene expression data](https://doi.org/10.1093/bioinformatics/btp616). Bioinformatics. 2010 Jan 1;26(1):139–40.

3. Langfelder P, Horvath S. [WGCNA: an R package for weighted correlation network analysis](https://doi.org/10.1186/1471-2105-9-559). BMC Bioinformatics. 2008 Dec 29;9(1):559.

4. Tzfadia O, Diels T, De Meyer S, Vandepoele K, Aharoni A, Van de Peer Y. [CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool](https://www.frontiersin.org/articles/10.3389/fpls.2015.01194). Frontiers in Plant Science. 2016;6.

5. Haider I, Yunmeng Z, White F, Li C, Incitti R, Alam I, et al. [Transcriptome analysis of the phosphate starvation response sheds light on strigolactone biosynthesis in rice](https://doi.org/10.1111/tpj.16140). The Plant Journal. 2023;114(2):355–70.

6. Wang Y, Duran HGS, van Haarst JC, Schijlen EGWM, Ruyter-Spira C, Medema MH, et al. [The role of strigolactones in P deficiency induced transcriptional changes in tomato roots](https://doi.org/10.1186/s12870-021-03124-0). BMC Plant Biol. 2021 Jul 23;21(1):349.

7. Saccenti E. What can go wrong when observations are not independently and identically distributed: a cautionary note on calculating correlations on combined data sets from different experiments or conditions. Frontiers in Systems Biology. 2023;3:3.

8. Geladi P, Kowalski BR. [Partial least-squares regression: a tutorial](https://doi.org/10.1016/0003-2670(86)80028-9). Analytica Chimica Acta. 1986 Jan 1;185:1–17.

9. Wold S, Ruhe A, Wold H, Dunn, III WJ. [The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses](https://doi.org/10.1137/0905052). SIAM J Sci and Stat Comput. 1984 Sep;5(3):735–43.

10. Searle SR. Linear Models. John Wiley & Sons; 1971. 565 p.

11. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers RJAN, van der Greef J, Timmerman ME. [ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data](https://doi.org/10.1093/bioinformatics/bti476). Bioinformatics. 2005 Jul 1;21(13):3043–8.

12. Thiel M, Féraud B, Govaerts B. [ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs](https://doi.org/10.1002/cem.2895). Journal of Chemometrics. 2017;31(6):e2895.

13. Auger S, Mournetas V, Chiapello H, Loux V, Langella P, Chatel JM. [Gene co-expression network analysis of the human gut commensal bacterium Faecalibacterium prausnitzii in R-Shiny](https://doi.org/10.1371/journal.pone.0271847). PLOS ONE. 2022 Nov 18;17(11):e0271847.

14. Ren F, Zhou Q, Meng Y, Guo W, Tang Q, Mei J. [RNA binding proteins are potential novel biomarkers of egg quality in yellow catfish](https://doi.org/10.1186/s12864-023-09220-9). BMC Genomics. 2023 Mar 16;24(1):121.

15. Nueda MJ, Conesa A, Westerhuis JA, Hoefsloot HCJ, Smilde AK, Talón M, et al. [Discovering gene expression patterns in time course microarray experiments by ANOVA–SCA](https://doi.org/10.1093/bioinformatics/btm251). Bioinformatics. 2007 Jul 15;23(14):1792–800.

16. Jarmund AH, Madssen TS, Giskeødegård GF. [ALASCA: An R package for longitudinal and cross-sectional analysis of multivariate data by ASCA-based methods](https://www.frontiersin.org/articles/10.3389/fmolb.2022.962431). Frontiers in Molecular Biosciences. 2022;9.

17. Guan J, Wang Y, Wang Y, Zhuang Y, Ji G. [SRGS: sparse partial least squares-based recursive gene selection for gene regulatory network inference](https://doi.org/10.1186/s12864-022-09020-7). BMC Genomics. 2022 Nov 30;23(1):782.

18. Nueda MJ, Sebastián P, Tarazona S, García-García F, Dopazo J, Ferrer A, et al. [Functional assessment of time course microarray data](https://doi.org/10.1186/1471-2105-10-S6-S9). BMC Bioinformatics. 2009 Jun 16;10(Suppl 6):S9.

19. Thissen U, Wopereis S, van den Berg SA, Bobeldijk I, Kleemann R, Kooistra T, et al. [Improving the analysis of designed studies by combining statistical modelling with study design information](https://doi.org/10.1186/1471-2105-10-52). BMC Bioinformatics. 2009 Feb 7;10(1):52.

20. de Jong S. [SIMPLS: An alternative approach to partial least squares regression](https://doi.org/10.1016/0169-7439(93)85002-X). Chemometrics and Intelligent Laboratory Systems. 1993 Mar 1;18(3):251–63.

21. Kvalheim OM. [Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots](https://doi.org/10.1002/cem.1289). Journal of Chemometrics. 2010;24(7-8):496–504.

22. Ito S, Yamagami D, Umehara M, Hanada A, Yoshida S, Sasaki Y, et al. [Regulation of Strigolactone Biosynthesis by Gibberellin Signaling](https://doi.org/10.1104/pp.17.00301). Plant Physiol. 2017 Jun;174(2):1250–9.

23. Marzec M. [Strigolactones and Gibberellins: A New Couple in the Phytohormone World?](https://doi.org/10.1016/j.tplants.2017.08.001) Trends in Plant Science. 2017 Oct 1;22(10):813–5.

24. Sun H, Guo X, Zhu X, Gu P, Zhang W, Tao W, et al. [Strigolactone and gibberellin signaling coordinately regulate metabolic adaptations to changes in nitrogen availability in rice](https://doi.org/10.1016/j.molp.2023.01.009). Mol Plant. 2023 Mar 6;16(3):588–98.

25. Farahbod M, Pavlidis P. [Untangling the effects of cellular composition on coexpression analysis](https://doi.org/10.1101/gr.256735.119). Genome Res. 2020 Jun 1;30(6):849–59.

26. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. [Weighted Gene Coexpression Network Analysis: State of the Art](https://doi.org/10.1080/10543400903572753). Journal of Biopharmaceutical Statistics. 2010 Mar 19;20(2):281–300.

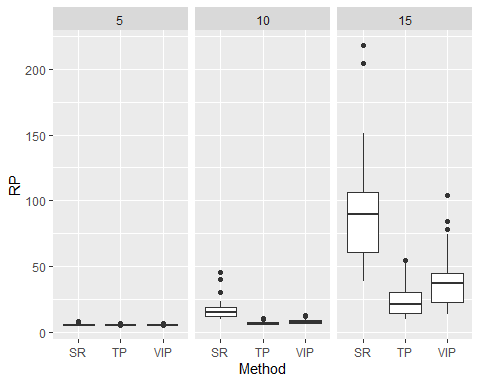
27. Chowdhury HA, Bhattacharyya DK, Kalita JK. [(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices](https://doi.org/10.1109/TCBB.2019.2893170). IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020 Jul;17(4):1154–73.

28. Rajalahti T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM. [Biomarker discovery in mass spectral profiles by means of selectivity ratio plot](https://doi.org/10.1016/j.chemolab.2008.08.004). Chemometrics and Intelligent Laboratory Systems. 2009 Jan 15;95(1):35–48.

29. Caldana C, Degenkolbe T, Cuadros-Inostroza A, Klie S, Sulpice R, Leisse A, et al. [High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions](https://doi.org/10.1111/j.1365-313X.2011.04640.x). The Plant Journal. 2011;67(5):869–84.

# Supplementary

Target projection in PLS2 models has not been described in the literature, we therefore tested our approach against the most common methods for feature selection in PLS namely variable importance in projection (VIP) and selectivity ratio (SR). Here we have similar but reduced simulation set up to the replicate test section; 3 levels of unstructured noise; matrix F-norm sizes (as indicated in the panels). In each of the 3 tests the is randomised 30 times resulting in different correlation structures between baits and spikes. Target projection using significantly outperforms the conventional approaches.



## Ranked correlations

Ranked correlations are calculated as:

Where *B* is the number of baits and is the ranked Pearson correlation coefficient between gene j and bait b in the context of non-bait genes.

## Fisher transformed Pearson correlation coefficient

To calculate Fisher transformed Pearson correlations (z) with Pearson’s r: