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

04 December, 2023

# Abstract

Experiments in plant transcriptomics are usually designed to induce variation in a pathway of interest (POI). Harsh experimental conditions can cause widespread transcriptional changes between groups. Discovering coexpression within a POI (here strigolactones) 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 POIs and related pathways (RPs) 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 POI or RPs. 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, or 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; Anders & Huber ([2010](#X6fa5e42e8bb074aaef1f9066b87c01d4c455181)); Robinson et al. ([2010](#ref-robinsonEdgeRBioconductorPackage2010))) or coexpression analysis (CoE; Langfelder & Horvath ([2008](#ref-langfelderWGCNAPackageWeighted2008)); Tzfadia et al. ([2016](#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 a partially characterized pathway. 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”).

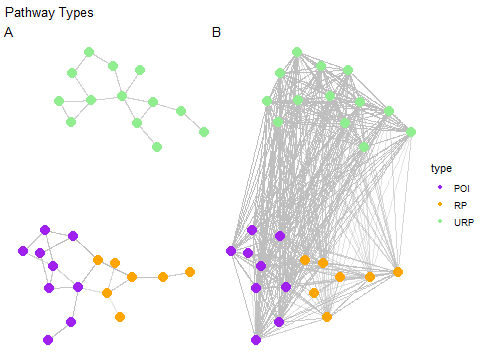


Figure 1: Pathway types exemplified, POI; pathway of interest, RP; related pathway, URP; unrelated pathway. A represents hypothetical biological pathways, B highlights the problem with detecting these pathways when dealing with data from designed experiments.

Figure 1 outlines pathway nomenclature that will be used throughout this text. *A* represents a true hypothetical pathway including a pathway of interest (POI), related pathway (RP) and an unrelated pathway (URP). All three types of pathway are assumed to be affected to varying degrees by the experimental design. 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 *B* from 1. Figure 2 exemplifies between-group and within-group driven correlations. 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 of both pairs of transcripts are high but the second pair’s correlation 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 one condition (facet *C*: P- condition in red) for the first gene pair. Because of this, 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 genes and/or pathways under investigation. 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 may be valuable. Variance between replicates is caused by non-controlled variations in plant development and environmental conditions. One replicate within a group may be slightly bigger than the others and therefore respond slightly different to a treatment, for example phosphate deficiency. A larger plant may start to suffer from the deficiency earlier than a smaller plant. This does not imply that the larger plant shows a different biological response to P deficiency. We can assume that the response is similar, but just occurs a bit faster. This in turn implies that the expression of the bait genes are induced a little bit faster than in the other replicates, but this will also be true for 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 this 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. For example in Haider et al. ([2023](#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 (Wang et al. ([2021](#ref-wangRoleStrigolactonesDeficiency2021))). Haider et al. ([2023](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) used ranked correlations, similar to MutRank (Poretsky & Huffaker ([2020](#X18618c4a7fe605fcd8d6b66c052f33c2f8fe63d))), 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 Saccenti ([2023](#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.

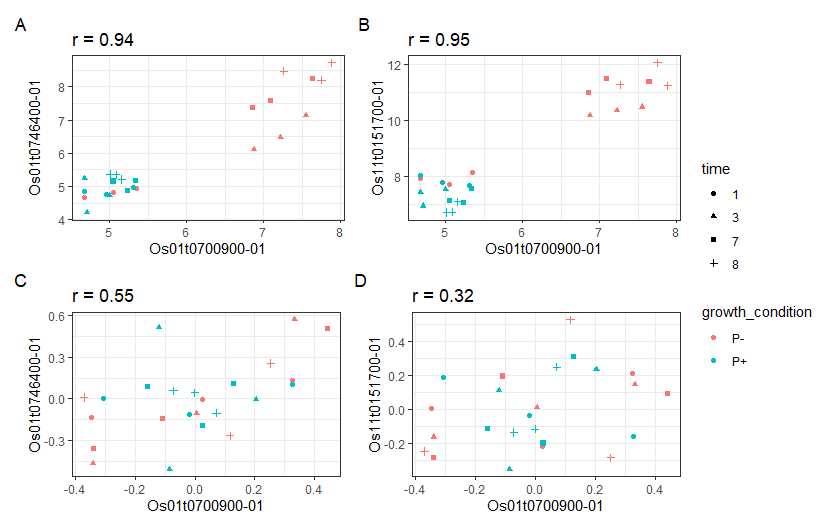


Figure 2: Correlation examples, Pearson’s correlation coefficient (r) indicated in facet titles. X axes, expression levels of CCD8 gene in rice, Y axes expression levels of example coexpression candidates. Top row A-B; *total CoE*, bottom row C-D *within-group CoE*. Gene pairs in both A and B have strong 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 which has been performed on the same genes in C-D.

For a successful coexpression analysis it is necessary to have a predefined set of baits that are already known to be involved in the mechanism under study. Methods like ranked correlations and partial least squares regression (PLS; Geladi & Kowalski ([1986](#Xd0e6c30a697db42abb2065af1a17da6eb418ea3)); Wold et al. ([1984](#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 i.e. bait 1 - candidate 1, bait 1 - candidate 2. 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; Searle ([1971](#ref-searleLinearModels1971))). ANOVA simultaneous component analysis (ASCA; Smilde et al. ([2005](#Xaca7e11f7e5f0a0465524ca94b5740f80e57dcd))) and its extension to imbalanced data via the generalised linear model (ASCA+; Thiel et al. ([2017](#ref-thielASCAAPCAExtensions2017))) have been developed to combine principal component analysis (PCA) with ANOVA and can therefore account for experimental design in dimensionality 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 (Auger et al. ([2022](#ref-augerGeneCoexpressionNetwork2022)), Ren et al. ([2023](#ref-renRNABindingProteins2023)), Nueda et al. ([2007](#ref-nuedaDiscoveringGeneExpression2007)), Jarmund et al. ([2022](#ref-jarmundALASCAPackageLongitudinal2022))) and CoE (Guan et al. ([2022](#ref-guanSRGSSparsePartial2022)), Nueda et al. ([2009](#ref-nuedaFunctionalAssessmentTime2009))). ASCA and PLS have been combined before in Thissen et al. ([2009](#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 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 Haider et al. ([2023](#Xc2b17888754b4d684e1893820af32b0e8e89a94))). Furthermore, this dataset is used as a guide to several simulation studies. 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 biosynthetic POI elucidation when a set of baits is already known and illustrate the use benefit of MASCARA, particularly in data coming from experiments designed for differential expression analysis.

# Materials and Methods

## RNAseq - Rice strigolactone dataset Haider et al. ([2023](#Xc2b17888754b4d684e1893820af32b0e8e89a94))

In order to guide the simulations we illustrate the characteristics of the simulated data 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 show 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.

Figure 3 shows the gene expression levels of 9 strigolactone pathway genes along with 9 highly differential genes that 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. The loading plot shows that although the 9 SL genes are mostly on the outside of the plot, many other genes are as well. Thus many differentially expressed genes are found important that are not part of the SL pathway, in other words ASCA does not detect spike SL genes from others that are similarly or more differentially expressed but not part of the SL pathway.

This marks the goal of this paper, which is to detect within-group coexpressed genes with a set of baits preferentially to other differentially expressed genes that are not part of the same pathway (with no within-group structure).

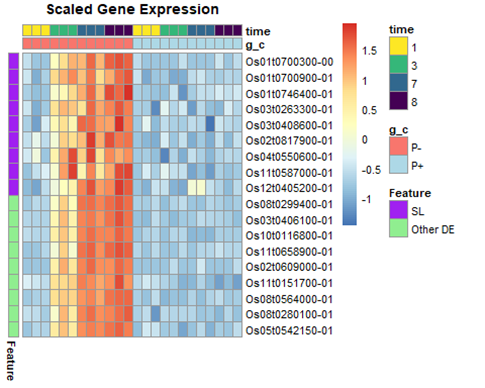


Figure 3: Real data overview. RNAseq of rice root. Heatmap of SL pathway and example highly differential genes.

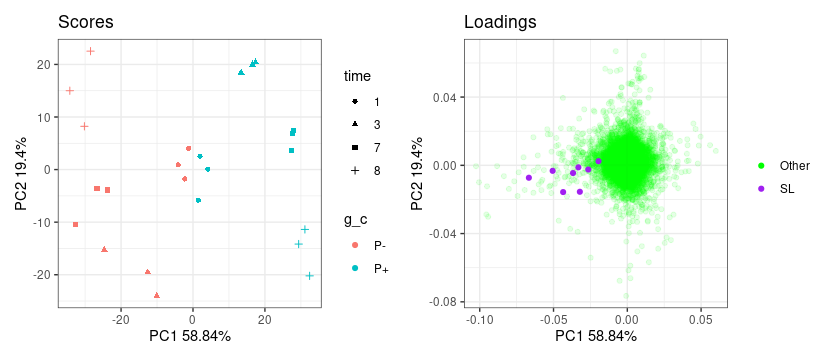


Figure 4: Scores and loadings of the ASCA model of the combined effect (condition + condition:time interaction).

## 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 and time. The simulated data is created by simulating score and loading values (as in Figure 5) for the combined effect () ) explaining the difference in time profiles of the P+ and P- condition and the main time effect , as well as, structured ambient variance (). This is to mimic variance due to gene expression as a function of growth condition and time as well as variance due to coexpression of genes in the same pathway on top of this experimentally induced DE variance.

## 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 non-pathway differential genes (genes with high between-group correlation but no within-group correlation). The expression levels of differential transcripts is modeled with the scores and loadings of the growth condition and its interaction with time ( and ). The disturbing time effect is modeled with scores and loadings ( and ). Also included is some random measurement noise, this is modeled as a matrix with normally distributed random values .

On top of the between-group effect there is an additional a correlation structure applied to the baits and spikes to create a model of coexpression that is independent to the variance induced by the experimental factors (). In this simulation we have a set of 4 baits that have positive loadings for the combined effect along with a variable number of other genes - “Combined Effect Responders”. For the structured ambient variance; 2 random sample offset vectors in (each column representing an effect caused by ambient conditions) which our baits and spikes contribute to (i.e. green and black bars in Figure 5 are non zero in ). 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 set thus consists of 24 samples and 2000 genes.

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. Loadings and are simulated to create the baits and spikes. As displayed in Figure 5, the green bar represents 4 baits (or known pathway genes), the black bar represents the 12 spikes (“unknown” genes of interest), yellow represents differentially expressed genes not in SL pathway, and purple; a set of differential genes not in SL pathway with an alternative treatment response.

In the baits, spikes and other DE genes 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. 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. represents random unstructured technical noise and the additional term in (1) represents a mean offset for each gene.

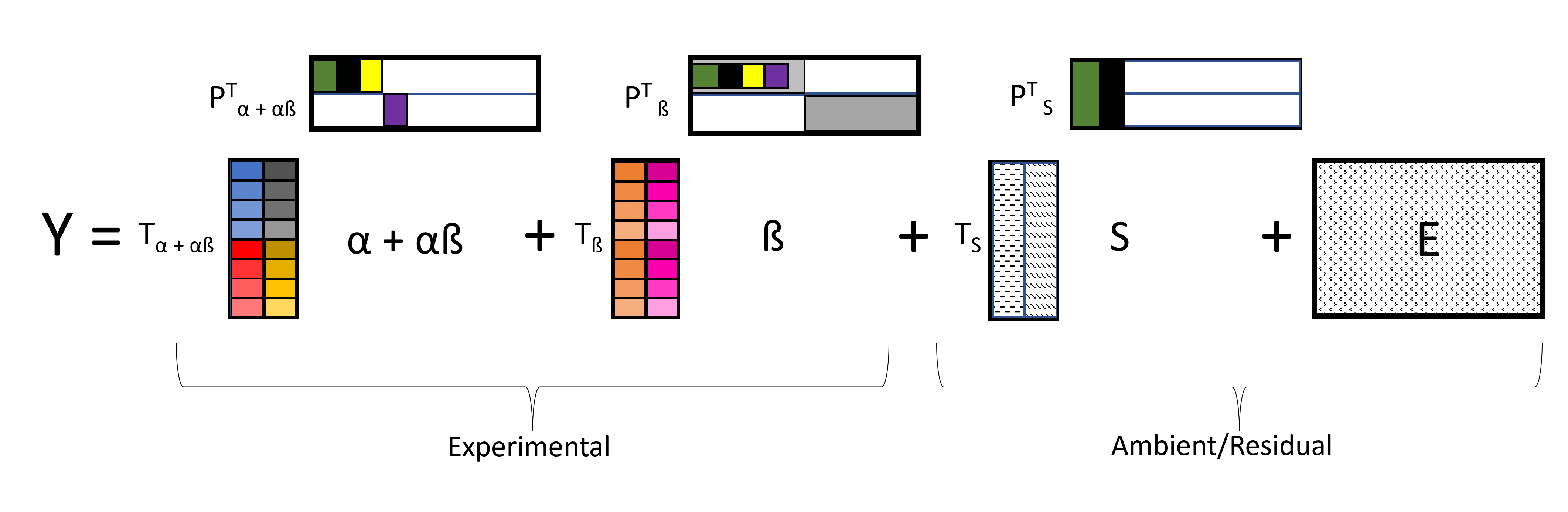


Figure 5: Coexpression simulation, simplified overview. Colours correspond to the same genes in the different loading vectors (*P*). Black; baits (known SL genes), green; spikes (unknown SL genes), yellow; other non-pathway DE genes, purple non-pathway DE genes with a different profile. In the score 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.

The issue to address is the presence of many non-pathway differential genes; those that have strong between group correlations but no within group correlation. Here the differential genes relate to the yellow bar in 5. We incrementally increase the number of non-zero values and decrease the number of 0 values, which in figure 5 relates to increasing the size of the yellow bar, as well as varying the F-norm of the combined effect matrix through the scalar parameter *d* below;

## Data simulation 2: Replication

In order to evaluate the generalisation of methods and indicate the levels of variance required for optimal performance we conduct a larger scale assay. Here we fix the number of non-pathway DE genes to 30 which corresponds (to the yellow bar size) and the size of the within correlation effect matrix (). We investigate a range of values for sizes of between variance (*d*) and random noise (*l*) across random structures (which correspond to repeated simulations of randomised values for matrices). Parameters were adjusted according to (3) on 20 randomised 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.

## Analysis and feature selection

### Ranked correlations

Pearson’s correlation coefficient is calculated between all columns of . For example the correlation between a SL gene and a non-SL gene is calculated as:

For ranked correlations *r* is calculated between each bait and all other genes, with the lowest rank for the highest correlation. Then for each bait the other genes are ranked based on *r*. These ranks are averaged across all baits, similarly to Poretsky & Huffaker ([2020](#X18618c4a7fe605fcd8d6b66c052f33c2f8fe63d)).

### ASCA+

ANOVA simultaneous component analysis (Smilde et al. ([2005](#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 Thiel et al. ([2017](#ref-thielASCAAPCAExtensions2017)). ASCA+ addresses the issue of biased effects matrices in unbalanced designs. The following is a summary of the algorithm description provided in Thiel et al. ([2017](#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 coded 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 Thiel et al. ([2017](#ref-thielASCAAPCAExtensions2017)).

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

### PLS regression - canonical mode

Partial Least Squares canonical mode (PLS2) is a multivariate regression technique used for modeling relationships between two sets of variables, and .

After splitting the data into the response matrix (containing only baits) (the green bar in the loadings) and the predictor matrix (containing the rest of the data) , resulting in the model of the data:

This model is estimated in our work 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 and respectively.
   2. Project and onto the first left and right singular vectors (, ) to obtain the scores and .
   3. Perform linear regression of and onto their corresponding scores and to obtain the respective loadings , and .
   4. Deflate and by removing the variance explained by the pair of scores (, ), and use the deflated matrices as inputs for the next iteration.

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.

### 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 Thiel et al. ([2017](#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 is used as input to PLS2 after being split into baits and remainder.

The genes are ranked based on target projection (Kvalheim ([2010](#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. The MASCARA pipeline in exemplified in Figure 6.

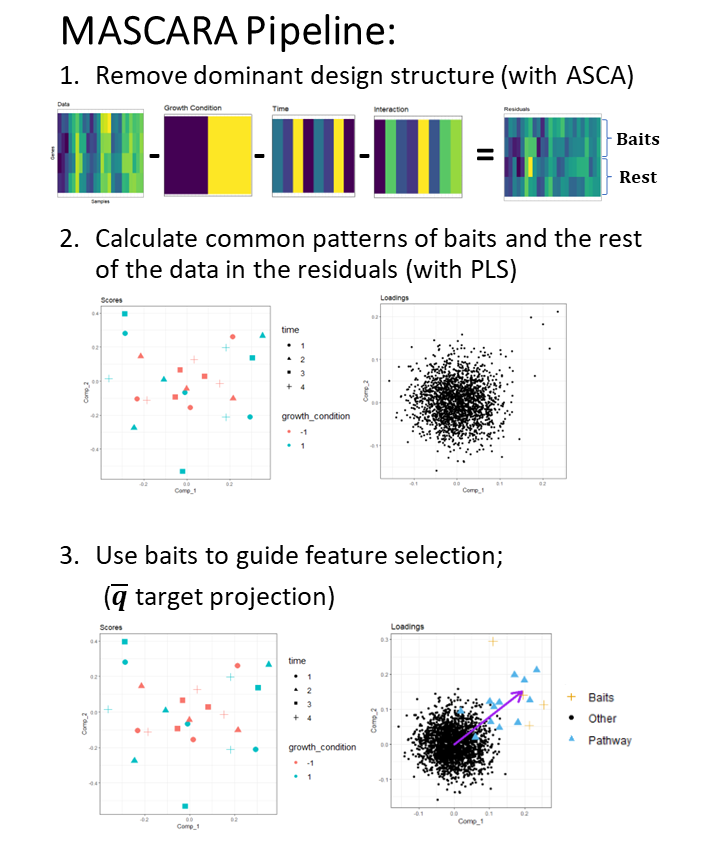


Figure 6: MASCARA pipeline overview

## Performance Evaluation

To evaluate performance we use log2 of the geometric mean rank (), defined as: for each spike *i* for n spikes.

For the real data, since the ground truth is not known we visualise relationships between baits and genes detected by the different methods rather than using a metric.

# 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 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 pathway 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 .

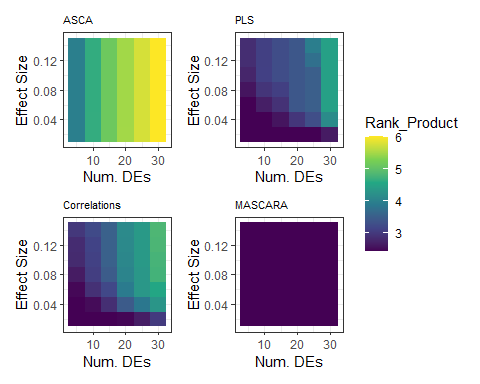


Figure 7: Results of simulation 1 for all 4 methods. The geometric mean rank as a function of the number of differentially expressed genes (x-axis) and combined effect size (y axis).

We can see in Figure 7 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 non-pathway DE genes. This was the expected result as the between group variance is removed with MASCARA.

## Simulation 2: Replication

The results from above were obtained from datasets with varying noise, with core aspects such as the underlying structure, the number of replicates, the strength of correlation between variables, and the size of the noise remained fixed. With the following, 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 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 above. Here we create 20 datasets with random combined effect and within group correlation structures.

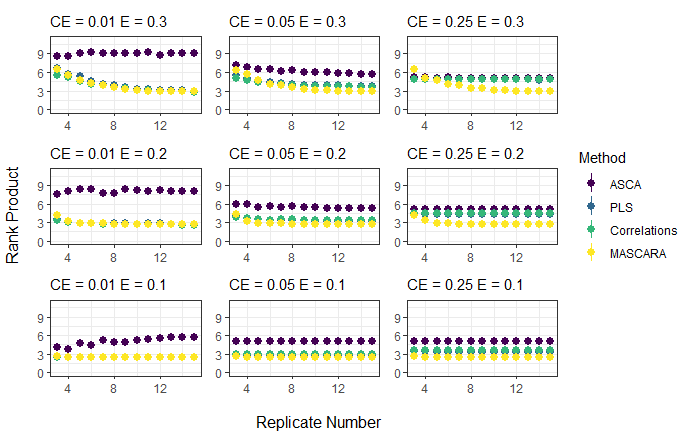


Figure 8: Simulation 3 results. Replications and variance type feasibility ranges. X axes; replicate number, y axes log2 transformed rank product. Each facet is created with set parameters for combined effect size () and random gaussian noise size (). Dots are median performance across 20 random structures.

It was expected that all geometric mean ranks converge towards a minimum with increasing numbers of replicates, see 8. Perfect detection of 12 spikes in top 12 candidates is achieved with a log2 geometric mean rank of 2.402955, MASCARA this was able to achieve this however the other methods were perturbed by the presence of the 30 non spike differential genes.With more replicates MASCARA always outperforms competitors. With higher combined effect size (CE) MASCARA outperforms the other methods but its performance is affected by random noise (E) size.

In simulation 2 we fixed the number of differential 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 7) higher numbers of DE genes results in higher geometric mean rank of the spikes for correlations, ASCA and PLS. MASCARA is designed specifically to mitigate the presence of large numbers of differential 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 pathway genes that is caused by independent non-controlled ambient factors.

Here we illustrate an application of MASCARA to the dataset from Haider et al. ([2023](#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 Haider et al. ([2023](#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 experiment independent relationship amongst the SL genes compared to their correlations with other DE genes, as displayed in Figure 9 where the blue distribution indicates the 9 transcripts from figures 3 and 4 and the red is the relationship between these 9 and other top differential genes.

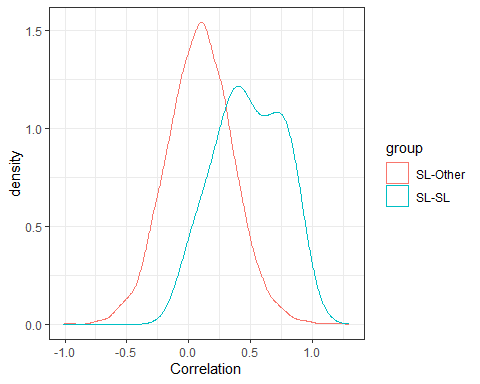


Figure 9: Distributions of Fisher transformed Pearson correlations in residuals from the ASCA model. Blue indicates correlations between SL pathway genes, red indicates correlations between SL pathway genes and top 10% (non-SL) DE genes; upregulated in P-

Some example top candidate genes are shown in Figure 10. The between versus within problem is again exemplified here. With several of the top candidates detected by correlations (ASCA and PLS, results not shown) showing 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 (Ito et al. ([2017](#Xbd25994ee0f44976988919f4c291e7808fe56a3)), Marzec ([2017](#Xea043edabf8aaad9d19b1e892525d9fffe3a696))). The interaction between these two pathways has been characterised recently under nitrogen limitation by Sun et al. ([2023](#X2e16474dbff84e9966c63a78f86e923d201544c)).

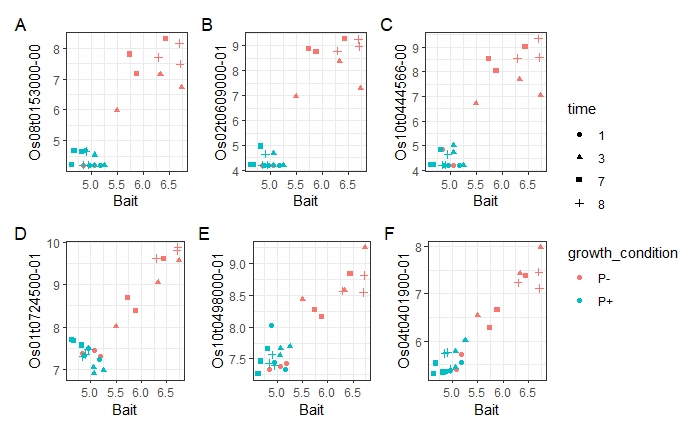


Figure 10: Selected candidates from A-C ranked correlations and D-F MASCARA. X axes; expression level of one of the bait genes (CCD8), y axes; expression levels of three example coexpression candidates for each method.

# 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 baits and the rest of the data. 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 plant growth, 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 Haider et al. ([2023](#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 (Smilde et al. ([2005](#Xaca7e11f7e5f0a0465524ca94b5740f80e57dcd)), Thiel et al. ([2017](#ref-thielASCAAPCAExtensions2017))) and PLS (Wold et al. ([1984](#ref-woldCollinearityProblemLinear1984))) are all affected by dominant between-group variance, which is a common occurrence in experimental data, exemplified here by the dataset from Haider et al. ([2023](#Xc2b17888754b4d684e1893820af32b0e8e89a94)) but discussed elsewhere (Farahbod & Pavlidis ([2020](#Xcc71cd2df4032d63494e28c6a7921129da1c090)), Zhao et al. ([2010](#ref-zhaoWeightedGeneCoexpression2010)), Chowdhury et al. ([2020](#X3e57ed0fe141b4b17e403de16a66af40da9b626))).

The feature selection stage of MASCARA uses an average vector constructed from the loadings of the baits onto which to project the loadings of other genes, which is similar to target projection in Kvalheim ([2010](#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 (Rajalahti et al. ([2009](#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. With the averaged target projection approach 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 Caldana et al. ([2011](#Xe7babcb509b3ef554dc77224d7bc0c8531e51b3)) where different combinations of experimental factors show differential within-group correlation structure. With MASCARA these could be missed, 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 mechanism of interest 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 off the structure of the Haider RNAseq dataset, however the method is applicable across other 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 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

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. FW thanks the Data Science Centre of the University of Amsterdam for partial financial support.

# References

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, 1–1. <https://doi.org/10.1038/npre.2010.4282.1>

Auger, S., Mournetas, V., Chiapello, H., Loux, V., Langella, P., & Chatel, J.-M. (2022). Gene co-expression network analysis of the human gut commensal bacterium Faecalibacterium prausnitzii in R-Shiny. *PLOS ONE*, *17*(11), e0271847. <https://doi.org/10.1371/journal.pone.0271847>

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., … Hannah, M. A. (2011). High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. *The Plant Journal*, *67*(5), 869–884. <https://doi.org/10.1111/j.1365-313X.2011.04640.x>

Chowdhury, H. A., Bhattacharyya, D. K., & Kalita, J. K. (2020). (Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(4), 1154–1173. <https://doi.org/10.1109/TCBB.2019.2893170>

Farahbod, M., & Pavlidis, P. (2020). Untangling the effects of cellular composition on coexpression analysis. *Genome Research*, *30*(6), 849–859. <https://doi.org/10.1101/gr.256735.119>

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1–17. <https://doi.org/10.1016/0003-2670(86)80028-9>

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

Haider, I., Yunmeng, Z., White, F., Li, C., Incitti, R., Alam, I., … Bouwmeester, H. J. (2023). Transcriptome analysis of the phosphate starvation response sheds light on strigolactone biosynthesis in rice. *The Plant Journal*, *114*(2), 355–370. <https://doi.org/10.1111/tpj.16140>

Ito, S., Yamagami, D., Umehara, M., Hanada, A., Yoshida, S., Sasaki, Y., … Asami, T. (2017). Regulation of Strigolactone Biosynthesis by Gibberellin Signaling. *Plant Physiology*, *174*(2), 1250–1259. <https://doi.org/10.1104/pp.17.00301>

Jarmund, A. H., Madssen, T. S., & Giskeødegård, G. F. (2022). ALASCA: An R package for longitudinal and cross-sectional analysis of multivariate data by ASCA-based methods. *Frontiers in Molecular Biosciences*, *9*. Retrieved from <https://www.frontiersin.org/articles/10.3389/fmolb.2022.962431>

Kvalheim, O. M. (2010). Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *Journal of Chemometrics*, *24*(7-8), 496–504. <https://doi.org/10.1002/cem.1289>

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. <https://doi.org/10.1186/1471-2105-9-559>

Marzec, M. (2017). Strigolactones and Gibberellins: A New Couple in the Phytohormone World? *Trends in Plant Science*, *22*(10), 813–815. <https://doi.org/10.1016/j.tplants.2017.08.001>

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C. J., Smilde, A. K., Talón, M., & Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA–SCA. *Bioinformatics*, *23*(14), 1792–1800. <https://doi.org/10.1093/bioinformatics/btm251>

Nueda, M. J., Sebastián, P., Tarazona, S., García-García, F., Dopazo, J., Ferrer, A., & Conesa, A. (2009). Functional assessment of time course microarray data. *BMC Bioinformatics*, *10*(Suppl 6), S9. <https://doi.org/10.1186/1471-2105-10-S6-S9>

Poretsky, E., & Huffaker, A. (2020). MutRank: an R shiny web-application for exploratory targeted mutual rank-based coexpression analyses integrated with user-provided supporting information. *PeerJ*, *8*, e10264. <https://doi.org/10.7717/peerj.10264>

Rajalahti, T., Arneberg, R., Berven, F. S., Myhr, K.-M., Ulvik, R. J., & Kvalheim, O. M. (2009). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems*, *95*(1), 35–48. <https://doi.org/10.1016/j.chemolab.2008.08.004>

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

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>

Saccenti, E. (2023). 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*, *3*, 3.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R.-J. A. N., van der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, *21*(13), 3043–3048. <https://doi.org/10.1093/bioinformatics/bti476>

Sun, H., Guo, X., Zhu, X., Gu, P., Zhang, W., Tao, W., … Zhang, Y. (2023). Strigolactone and gibberellin signaling coordinately regulate metabolic adaptations to changes in nitrogen availability in rice. *Molecular Plant*, *16*(3), 588–598. <https://doi.org/10.1016/j.molp.2023.01.009>

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

Thissen, U., Wopereis, S., van den Berg, S. A., Bobeldijk, I., Kleemann, R., Kooistra, T., … Smilde, A. K. (2009). Improving the analysis of designed studies by combining statistical modelling with study design information. *BMC Bioinformatics*, *10*(1), 52. <https://doi.org/10.1186/1471-2105-10-52>

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

Wang, Y., Duran, H. G. S., van Haarst, J. C., Schijlen, E. G. W. M., Ruyter-Spira, C., Medema, M. H., … Bouwmeester, H. J. (2021). The role of strigolactones in P deficiency induced transcriptional changes in tomato roots. *BMC Plant Biology*, *21*(1), 349. <https://doi.org/10.1186/s12870-021-03124-0>

Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743. <https://doi.org/10.1137/0905052>

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

# 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.

