JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 0, 2019

Pp. 1–18
DOI: 10.1089/cmb.2019.0221

Research Article

A Genetic Algorithm to Optimize Weighted Gene Co-Expression Network Analysis

DAVID TOUBIANA,[1] RAMI PUZIS,[2] AVI SADKA,[3] and EDUARDO BLUMWALD[1]

ABSTRACT

Weighted gene co-expression network analysis (WGCNA) is a widely used software tool that is used to establish relationships between phenotypic traits and gene expression data. It generates gene modules and then correlates their first principal component to phenotypic traits, proposing a functional relationship expressed by the correlation coefficient. However, gene modules often contain thousands of genes of different functional backgrounds. Here, we developed a stochastic optimization algorithm, known as genetic algorithm (GA), optimizing the trait to gene module relationship by gradually increasing the correlation between the trait and a subset of genes of the gene module. We exemplified the GA on a Japanese plum hormone profile and an RNA-seq dataset. The correlation between the subset of module genes and the trait increased, whereas the number of correlated genes became sufficiently small, allowing for their individual assessment. Gene ontology (GO) term enrichment analysis of the gene sets identified by the GA showed an increase in specificity of the GO terms associated with fruit hormone balance as compared with the GO enrichment analysis of the gene modules generated by WGCNA and other methods.

Keywords: genetic algorithm, plant hormones, *Prunus salicina*, Japanese plum, RNAseq, weighted gene co-expression network analysis.

1. INTRODUCTION

With the advent of next-generation sequencing (NGS), transcript quantification has become possible for virtually all living organisms. As a consequence of steep reductions of per-base costs and progressive technological enhancements (Wadapurkar and Vyas, 2018), the number of studies employing NGS for transcript quantification has increased steadily (Lachmann et al., 2018). One of the aims of NGS studies is the detection of gene clusters that change their expression patterns in a co-ordinated manner throughout different conditions. To do so, a software tool, coined weighted gene co-expression network analysis (WGCNA) has been developed (Langfelder and Horvath, 2008). WGCNA generates correlation networks based on gene expression data. Highly interconnected genes are then clustered into modules, based on a topological overlap measure (Langfelder and Horvath, 2008). Subsequently, the first principal components

[1]Department of Plant Sciences, University of California, Davis, Davis, California.
[2]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel.
[3]Department of Fruit Tree Sciences, ARO, The Volcani Center, Rishon LeZion, Israel.

of modules are computed, termed module eigengenes (ME), which then can be correlated to trait data. Based on the correlation coefficient between MEs and the external traits, modules that likely include genes significantly impacting the trait of study are identified. WGCNA is a method that has been extensively used, for example, for the proposal of marker genes associated with Alzheimer's disease (Miller et al., 2010) or to highlight conservation and divergence of gene expression between human and chimpanzee brains (Oldham et al., 2006).

Modern software for the alignment of RNA-seq data to existing genomes (Dobin et al., 2013), coupled with high-performance computing, has the capability to rapidly align transcripts up to the coding sequence (CDS) level, rendering datasets of quantified expression (Anders et al., 2015) with hundreds of thousands of features. Thus, WGCNA often generates modules containing thousands of different CDSs. Although their corresponding MEs can still be correlated to particular traits, the number of CDSs per module is overwhelming, exacerbating the identification of marker genes or the comparison between networks. Aiming at reducing the number of genes in a module, we developed a genetic algorithm (GA).

A GA is a heuristic search method that is capable of detecting a near-optimal solution to a given search problem via the application of the principles of the theory of evolution (Mitchell, 1996); that is, the GA gradually increases the quality (*fitness*) of a collection of solutions over a given number (g) of *generations.* A GA initiates the search with a random population of x solutions, followed by *natural selection*, whereby a *fitness* value is assigned to each solution, increasing or decreasing their chances to reproduce with another ''solution'' (*individual*) in the population. The *reproduction* process is characterized by one or more *recombination* events, generating the next generation of solutions (*offspring*). Finally, a *mutation* process may occur modifying the ''genetic'' makeup of the *offspring*. GAs are commonly used for the analyses of biological data, for example, for multiple sequence alignments (Notredame and Higgins, 1996; Gondro and Kinghorn, 2007), motif discovery (Wong et al., 2011), RNA structure prediction (Vanbatenburg et al., 1995), etc.

Here, we describe the development of a GA customized for the optimization of CDSs/genes of modules in respect to a phenotypic trait. The GA successfully narrowed down the set of potential candidate genes to a number sufficiently small allowing for their individual analysis and concomitantly increased the trait to gene specificity. We applied our GA on datasets comprising RNA-seq and hormone-associated metabolites from two varieties of Japanese plum fruit during on-the-tree ripening and during postharvest storage (Farcuh et al., 2017, 2018, 2019). The two cultivars differed in their ripening behavior (Kim et al., 2015): a climacteric cultivar Santa-Rosa (SR), producing increased levels of autocatalytic ethylene and respiration rates during fruit ripening, and its nonclimacteric bud sport mutant Sweet Miriam (SM) showing no ethylene production or high respiration rates during ripening.

2. METHODS

2.1. Datasets acquisition and processing

Datasets for RNA-seq and hormone levels were adopted from Farcuh et al. (2017, 2018, 2019). Preprocessing and quantification of transcripts and hormones were performed as described therein.

2.2. Weighted gene co-expression network analysis settings

All data were log-transformed for usage with WGCNA. WGCNA version 1.60 was used. Execution of WGCNA was performed as instructed in Langfelder and Horvath (2008) and the associated tutorial. Soft thresholding of the adjacency matrix was achieved at $\beta = 17$. The manual hybrid model (Fig. S4) was chosen to generate gene modules.

2.3. Genetic algorithm parameterization

Before the initiation of the GA, all negatively correlated genes to each trait were removed from the module expression matrix. The GA was run with all parameters set to their respective default values. The GA was set to achieve an absolute $|fitness|$ value.

2.4. Gene Ontology term enrichment analysis

For the GO term enrichment analysis, we used Agrigo (http://bioinfo.cau.edu.cn/agriGO/) (Du et al., 2010; Tian et al., 2017). We performed a singular enrichment analysis against the Phytozome v11.0 *P. persica* v.2.1 genome.

2.5. Availability and dependencies

The source code for the GA is available under: https://github.com/toubiana/GENETIC_ALGORITHM
Operating system(s): Platform independent
Programming language: R
The GA operates in R without any dependencies, but we recommend installing package WGCNA for gene module generation.

3. RESULTS

3.1. Hormones specifically correlated to two module eigengenes

This study is based on datasets of Japanese plums previously published (Farcuh et al., 2017, 2018, 2019). In brief, gene expression and fruit hormone contents were measured in two cultivars: a climacteric SR and its nonclimacteric bud-sport mutant SM, during fruit development on-the-tree as well as throughout postharvest storage and in response to ethylene treatments. Gene expression analyses and contents of abscisic acid (ABA), ethylene, indole-3-acetic acid (IAA), gibberellins (GA_1 and GA_3), salicylic acid (SA), and the cytokinins trans-zeatin (tZ) and its precursor isopentenyl (iP) were performed at 12 different fruit development stages (Farcuh et al., 2017, 2018, 2019).

Transcripts from RNA-seq sequencing were aligned and quantified at the CDS level against the *Prunus persica* genome (Verde et al., 2013, 2017). All 12 storage stages were compared for differentially expressed CDSs (Anders and Huber, 2010), and a total of 18,714 differentially expressed CDSs were identified. Subsequently, the relative expression values of all 18,714 CDSs were fed into the WGCNA pipeline (Methods section). Overall, 18,621 out of the 18,714 transcripts were clustered into 14 modules (Table 1), whereas 93 CDSs remained unclustered. MEs were computed for all 14 modules and correlated to the profiles of all 8 hormones. The correlation analysis revealed that all hormones had the strongest positive correlation to either ME darkslateblue, where the corresponding module contained 2142 CDSs, or ME turquoise, where the corresponding module contained 4437 CDSs (Fig. 1). The following correlation coefficients were recorded between ME darkslateblue and ABA = 0.64, ethylene = 0.79, and IAA = 0.77; and ME turquoise and $GA_1 = 0.73$, $GA_3 = 0.79$, iP = 0.36, SA = 0.94, and tZ = 0.96, respectively (Fig. 1).

3.2. Gene Ontology enrichment analysis of modules darkslateblue and turquoise

The Gene Ontology (GO) initiative was developed to provide a system, in which sets of genes can be classified hierarchically in a graph-like structure (Harris et al., 2008). GO term enrichment analysis is the natural successor to WGCNA, whereby genes of modules identified via correlation analysis are analyzed to establish a functional relationship between the genes and the trait under investigation.

Table 1. Weighted Gene Co-expression Network Analysis Modules

Module	*Number of CDSs*	*Corresponding number of genes*
Module antiquewhite4	1819	1451
Module blue	4465	3449
Module coral1	2047	1521
Module darkslateblue	**2142**	**2142**
Module greenyellow	85	77
Module gray60	535	463
Module lightcyan	41	34
Module lightyellow	1651	1245
Module magenta	486	414
Module midnightblue	52	49
Module pink	121	93
Module purple	155	125
Module royalblue	585	533
Module turquoise	**4437**	**3239**

CDS, coding sequence.
Bold values correspond to the modules with the strongest correlations to hormones.

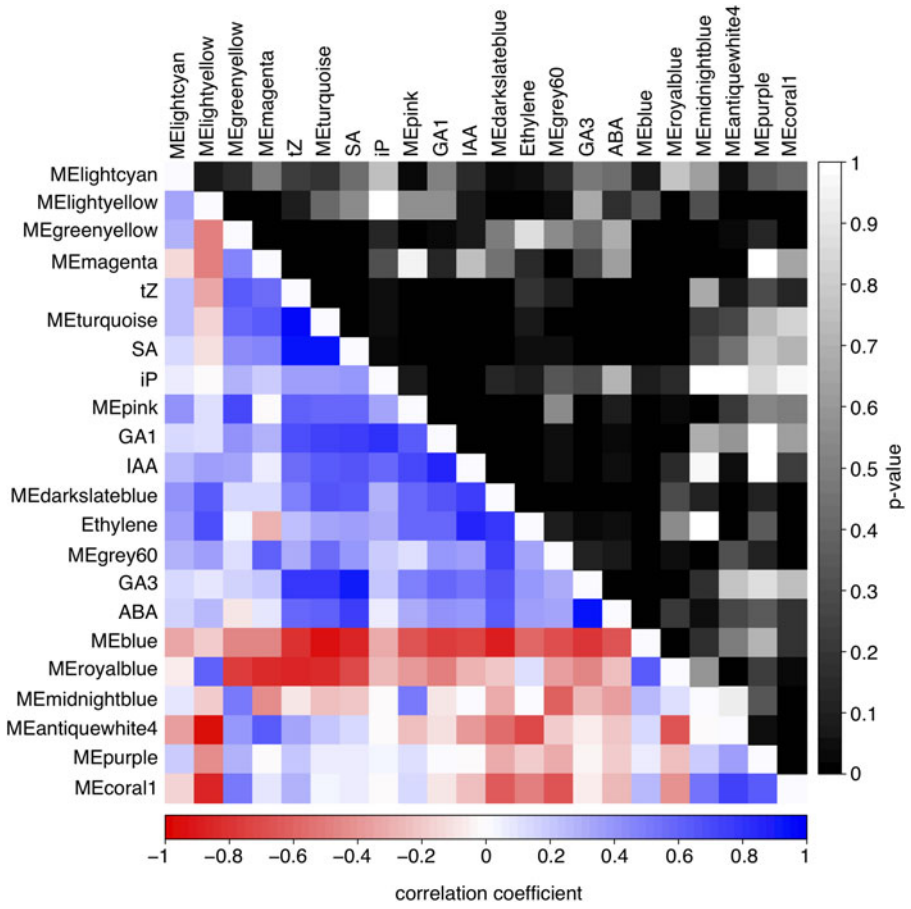


FIG. 1. ME to hormones correlation heatmap. Heatmap representation of the correlation analysis of 14 MEs and 8 hormones. The lower triangle illustrates the correlation coefficients, where a red rectangle represents a negative correlation and a blue rectangle a positive correlation. The upper triangle (shaded area) represents the corresponding *p*-values. Variables on the *x* and *y* axes are ordered as determined by hierarchical clustering. ME, module eigengenes.

We identified the corresponding genes to all the CDSs within the modules darkslateblue and turquoise (Supplementary Data S1) and performed GO enrichment analysis (Tables 2 and 3). Genes of module darkslateblue were categorized into 67 different significant (determined by a false discovery rate multiple hypothesis testing correction) GO terms, whereas genes of module turquoise were categorized into 58 significant GO terms. In total, 110 different GO terms were represented in both modules.

3.3. Narrowing candidate genes through the application of the genetic algorithm

To optimize the correlation between hormones and the genes within a module, and to reduce the number of candidate genes, we developed a GA for trait-related gene selection (TRGS). We denoted the matrix of a gene expression module as *Expr* of size $m \times n$, where $1 \leq i \leq m$ represented conditions (*Con*) and $1 \leq j \leq n$ genes (*Gen*). Within a given *Expr*, the TRGS algorithm identified a subset of genes showing the greatest correlation coefficient of its first principal component *E* to a given hormone *H*. Given *Expr*, the TRGS optimization objective can be defined as:

$$ARGMAX_C\{cor(prcomp(Expr[\bullet, C]), H)\}$$

where *C* is a subset of genes in the module, *Expr*[•, C] is the expression matrix narrowed down to *C*, *prcomp* is the first principal component of a matrix, and *cor* is the Pearson correlation.

A typical GA defines the population (structure and initial set of possible solutions), recombination and mutation (operators used to search through the space of possible solutions), and the fitness function (the optimization objective). Following the bio-inspired terminology common to GAs, we will refer to solutions

Table 2. Significant Gene Ontology Terms Associated with Module Darkslateblue

Term	p	*FDR*
Intracellular part	0.000012	0.001
Cell part	0.000017	0.001
Cytoplasm	0.000012	0.001
Cell	0.000017	0.001
Intracellular	0.0000055	0.001
Purine ribonucleoside monophosphate metabolic process	0.0000056	0.0011
Ribonucleoside monophosphate metabolic process	0.0000056	0.0011
Ribonucleoside triphosphate metabolic process	0.0000056	0.0011
Purine ribonucleoside triphosphate metabolic process	0.0000056	0.0011
ATP metabolic process	0.0000054	0.0011
Small molecule metabolic process	0.0000021	0.0011
Nucleoside triphosphate metabolic process	0.0000067	0.0011
Purine nucleoside triphosphate metabolic process	0.0000056	0.0011
Nucleoside monophosphate metabolic process	0.0000067	0.0011
Single-organism cellular process	0.0000065	0.0011
Purine nucleoside monophosphate metabolic process	0.0000056	0.0011
Organophosphate metabolic process	0.000011	0.0015
Purine ribonucleotide metabolic process	0.000012	0.0015
Ribonucleotide metabolic process	0.000012	0.0015
Purine ribonucleoside metabolic process	0.000014	0.0016
Purine nucleoside metabolic process	0.000014	0.0016
Purine nucleotide metabolic process	0.000021	0.002
Purine-containing compound metabolic process	0.000024	0.002
Nucleoside metabolic process	0.000025	0.002
Ribonucleoside metabolic process	0.000023	0.002
Ribose phosphate metabolic process	0.000024	0.002
Glycosyl compound metabolic process	0.000025	0.002
Organic acid metabolic process	0.000065	0.005
Nucleoside phosphate metabolic process	0.000089	0.0063
Catabolic process	0.000088	0.0063
Oxoacid metabolic process	0.0001	0.0068
Nucleobase-containing small molecule metabolic process	0.00013	0.0086
Nucleotide metabolic process	0.00016	0.01
Organic substance catabolic process	0.00019	0.012
Single-organism catabolic process	0.00023	0.013
Generation of precursor metabolites and energy	0.00023	0.013
Cytoplasmic part	0.00032	0.016
Organonitrogen compound metabolic process	0.0003	0.017
Organelle membrane	0.00054	0.024
Nucleoside diphosphate metabolic process	0.00047	0.025
Intracellular membrane-bounded organelle	0.0012	0.034
Intracellular organelle	0.0014	0.034
Membrane-bounded organelle	0.0012	0.034
Organelle	0.0015	0.034
Endomembrane system	0.0013	0.034
Macromolecular complex	0.00095	0.034
Establishment of localization	0.00071	0.037
Protein localization to membrane	0.0008	0.039
Establishment of protein localization to membrane	0.0008	0.039
Localization	0.00085	0.04
ADP metabolic process	0.0011	0.041
Establishment of protein localization	0.0012	0.041
Purine ribonucleoside diphosphate metabolic process	0.0011	0.041
Single-organism membrane organization	0.001	0.041

(*continued*)

TABLE 2. (CONTINUED)

Term	p	*FDR*
Nucleoside diphosphate phosphorylation	0.00093	0.041
Single-organism carbohydrate metabolic process	0.0012	0.041
Single-organism carbohydrate catabolic process	0.0011	0.041
ATP generation from ADP	0.0011	0.041
Nucleotide phosphorylation	0.00093	0.041
Purine nucleoside diphosphate metabolic process	0.0011	0.041
Ribonucleoside diphosphate metabolic process	0.0011	0.041
Carboxylic acid metabolic process	0.00094	0.041
Glycolytic process	0.0011	0.041
Mitochondrial inner membrane	0.0019	0.041
Organelle inner membrane	0.0023	0.043
Mitochondrial membrane	0.0021	0.043
Carbohydrate catabolic process	0.0013	0.046

FDR, false discovery rate.

of the optimization problem as *chromosomes*. An overview of TRGS can be seen in Figure 2. TRGS main components are discussed next.

3.4. Trait-related gene selection initial population

TRGS intakes seven arguments, namely: (1) the expression matrix *Expr*; (2) the population size *ps*, corresponding to the number of chromosomes within a single generation. The default value of *ps* was 1000 and remained unchanged throughout the algorithm execution. Each chromosome was represented by a binary vector of size *n*, where the value ''1'' represented a ''selected gene'' and ''0'' represented an ''ignored gene''; (3) number of genes *ng*, representing the number of genes that should be available for chromosomes in the initial generation (the default is set to 10); (4) crossover events *ce*, which determined the number of crossover events that should occur during recombination of two chromosomes (the default is set to 1); (5) mutation rate *mr*, which determined the chance of each gene in a chromosome to flip its value reciprocally (the default is set to 0.001%); (6) number of generations *g*, which specified the number of generations TRGS should be running (the default is set to 1000); and (7) the trait to which the subset of genes should be correlated. As a result, the initial generation was created with chromosomes chrom_1, $\text{chrom}_2 \ldots \text{chrom}_k \ldots$ of length *n*, where $1 \leq k \leq ps$ and where *ng* random cells of each chromosome were set to 1 and the rest were set to 0.

3.5. Trait-related gene selection fitness function

The fitness value of each chromosome determined its chances to participate in a reproduction event. Here, we first selected a subset of columns of the expression matrix corresponding to the selected genes in each chromosome, for example, $Expr[\bullet, Chrom_1]$, $Expr[\bullet, Chrom_2] \ldots Expr[\bullet, Chrom_k]$ (Fig. 2a). The fitness value was defined as the correlation of the first principal component E of $Expr[\bullet, Chrom_k]$ to the trait parameter (Fig. 2b):

$$E_k = prcomp(Expr_m[\bullet, Chrom_k])$$

$$Fitness\ (k) = cor(E_k, H)$$

3.6. Trait-related gene selection recombination

Chromosomes with greater fitness have higher chances to participate in a recombination event to generate chromosomes for the subsequent generation. The GA considered that recombination took place between two different chromosomes (mother and father chromosome) only. The amount of crossover events between mother and father chromosomes was determined by parameter *ce*. The locus *l* for the crossover event was determined randomly. At the specified *l*, the mother and father chromosomes were

Table 3. Significant Gene Ontology Terms Associated with Module Turquoise

Term	*term_type*	p	*FDR*
Cytoplasm	C	3.6E-26	1.2E-23
Cytoplasmic part	C	2.3E-26	1.2E-23
Intracellular	C	1.2E-24	2.5E-22
Cell part	C	4E-22	5.2E-20
Cell	C	4E-22	5.2E-20
Intracellular part	C	1.4E-21	1.5E-19
Translation	P	5.2E-22	4.5E-19
Peptide metabolic process	P	2.6E-22	4.5E-19
Cellular amide metabolic process	P	6.7E-22	4.5E-19
Peptide biosynthetic process	P	9.3E-22	4.5E-19
Amide biosynthetic process	P	9.3E-22	4.5E-19
Structural molecule activity	F	2.1E-21	2.4E-18
Ribosome	C	1.6E-19	1.3E-17
Macromolecular complex	C	1.6E-19	1.3E-17
Ribonucleoprotein complex	C	5.3E-19	3.4E-17
Intracellular ribonucleoprotein complex	C	5.3E-19	3.4E-17
Intracellular organelle	C	1.4E-18	8E-17
Organelle	C	1.5E-18	8.1E-17
Structural constituent of ribosome	F	1.9E-19	1.1E-16
Organonitrogen compound metabolic process	P	1.7E-18	6.7E-16
Organonitrogen compound biosynthetic process	P	3.5E-18	1.2E-15
Intracellular nonmembrane-bounded organelle	C	3.6E-15	1.7E-13
Nonmembrane-bounded organelle	C	3.6E-15	1.7E-13
Endomembrane system	C	0.000000011	0.00000049
Nitrogen compound metabolic process	P	5.2E-09	0.0000016
Cellular nitrogen compound metabolic process	P	6.9E-09	0.0000019
Macromolecule biosynthetic process	P	0.000000012	0.000003
Cellular macromolecule biosynthetic process	P	0.000000017	0.0000037
Intracellular organelle part	C	0.00000037	0.000015
organelle part	C	0.0000004	0.000015
Biosynthetic process	P	0.00000012	0.000024
Organic substance biosynthetic process	P	0.0000003	0.000055
Cellular biosynthetic process	P	0.00000037	0.00006
Gene expression	P	0.00000035	0.00006
Cellular nitrogen compound biosynthetic process	P	0.0000006	0.000092
Protein folding	P	0.0000022	0.00031
Endoplasmic reticulum	C	0.0000092	0.00033
Membrane-bounded organelle	C	0.000016	0.00052
Intracellular membrane-bounded organelle	C	0.000016	0.00052
Protein complex	C	0.000048	0.0015
Organelle membrane	C	0.0001	0.003
Envelope	C	0.00013	0.0034
Organelle envelope	C	0.00013	0.0034
Establishment of protein localization	P	0.00004	0.0054
Protein transport	P	0.000059	0.0076
Protein localization	P	0.000084	0.01
Cellular process	P	0.00012	0.014
Whole membrane	C	0.00055	0.014
Mitochondrial part	C	0.00062	0.015
RNA binding	F	0.000044	0.017
Bounding membrane of organelle	C	0.00082	0.02
Membrane protein complex	C	0.00097	0.022
Vesicle-mediated transport	P	0.00036	0.04
Macromolecule localization	P	0.00042	0.045

(*continued*)

TABLE 3. (CONTINUED)

Term	*term_type*	p	*FDR*
Mitochondrial envelope	C	0.002	0.045
Golgi apparatus part	C	0.0021	0.046
Guanyl ribonucleotide binding	F	0.00021	0.049
GTP binding	F	0.00021	0.049

split, so that the child chromosome $chrom_{offspring}$ was composed of $chrom_{mother}[1{:}l]$ and $chrom_{father}[(l+1){:}n]$ (Fig. 2c). Each offspring generation was composed of the same number of chromosomes corresponding to *ps.*

3.7. Trait-related gene selection mutation

After the recombination step, each $chrom_{offspring}$ was subjected to a possible mutation event, where each value in the chromosome could change from $0 \rightarrow 1$ or from $1 \rightarrow 0$, respectively, determined by *mr* (Fig. 2d).

3.8. Trait-related gene selection return values

After the TRGS had run for *g* generations, it returned a set of binary vectors corresponding to the final generation (denoted *lastpopulation*) and the fitness values of each one of them. We also recorded the average fitness of every generation for further analysis of the algorithm performance.

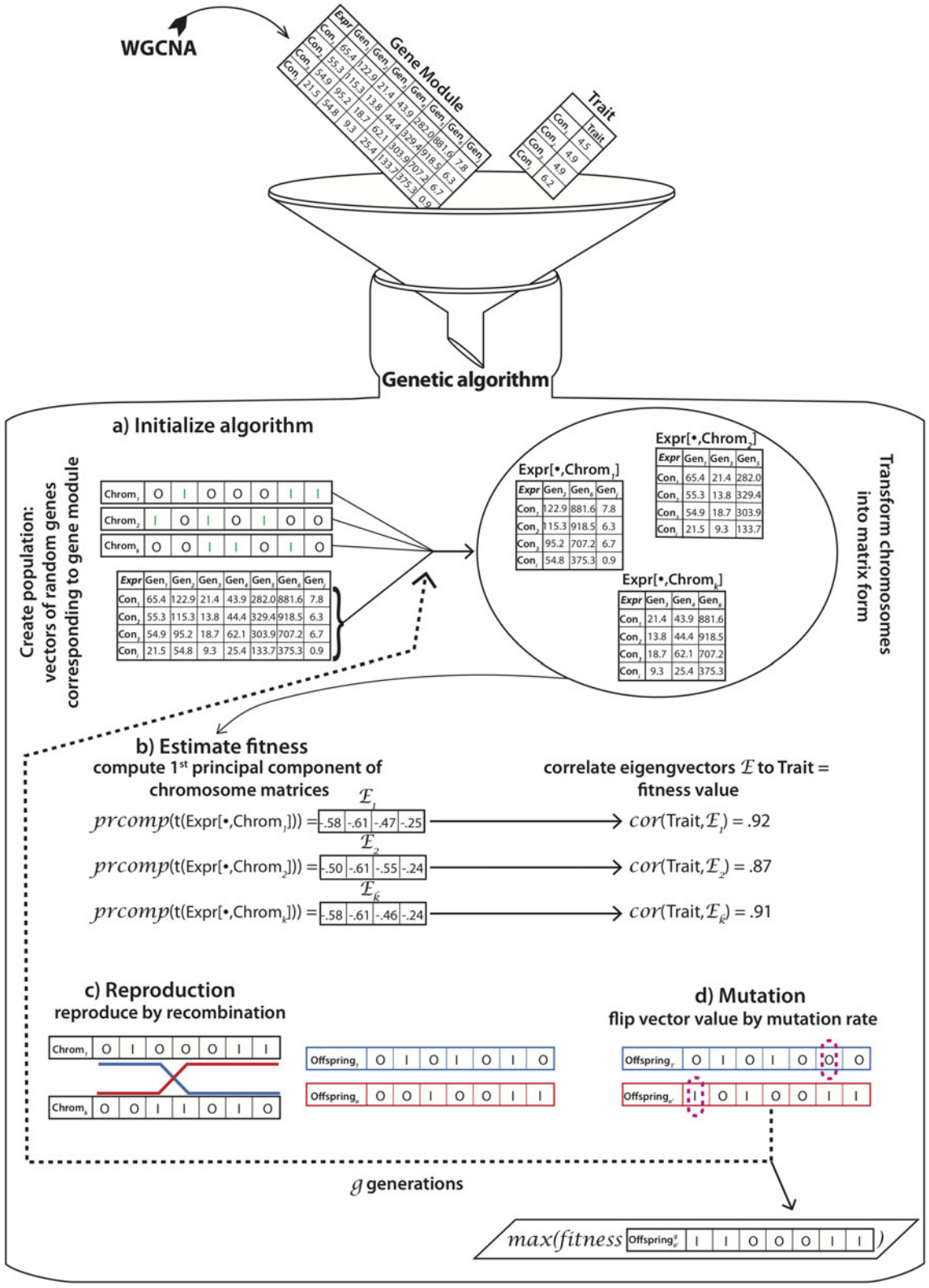
3.9. Genetic algorithm performance evaluation

We employed the GA on module darkslateblue for hormones ABA, ethylene, and IAA and on module turquoise for hormones GA_1, GA_3, iP, SA, and tZ. To test for its performance, the GA was executed 100 times (iterations) for each hormone to its respective module. We recorded the average correlation for each generation across the 100 iterations and its respective standard deviations (Fig. 3). Unequivocally, the GA increased the magnitude of correlation over 1000 generations in all 100 iterations. For the final generation, ABA recorded an average correlation coefficient of 0.97, when correlating its *E* to a subset of CDSs of module darkslateblue versus the original value of 0.64 to ME darkslateblue. Further, the greatest correlation between any CDS in module darkslateblue to ABA was estimated at 0.91 (Table 4). The respective values for ethylene were 0.98 versus 0.79 and 0.9; for IAA, 0.98 versus 0.77 and 0.98; for ME turquoise, the respective values for GA_1 were 0.94 versus 0.73 and 0.96; for GA_3, 0.98 versus 0.79 and 0.96; iP increased to 0.91 versus 0.36 and 0.62; SA recorded 0.98 versus 0.94 and 0.92; and tZ increased to 0.99 versus 0.96 and 0.95. Other GA-associated performance measures are summarized in Supplementary Figures S1–S3.

3.10. Threshold settings for gene selection

To select for putatively biologically meaningful genes, threshold settings were applied on the optimized solutions. Given that the application of the GA on different datasets produces different outcomes, the

FIG. 2. GA overview. A gene expression dataset *Expr* corresponding to the subset of genes of the gene module generated by WGCNA of size $m \times n$, where $1 \leq i \leq m$ represents conditions (*Con*) and $1 \leq j \leq n$ genes (*Gen*) and a dataset of traits *T* is fed into the GA; **(a)** Algorithm initialization: chromosomes $Chrom_1$, $Chrom_2$…$Chrom_k$ are produced as binary vectors of size *m*, where the value ''1'' represents a ''selected gene'' and 0 represents an ''ignored gene.'' Subsequently, each chromosome is transformed into matrix form, such that $Expr[\bullet Chrom_k]$ where $Chrom_k$ is a subset of genes in the module; **(b)** Fitness computation: the fitness value is determined by computing the first principal component of transposed $Expr[\bullet Chrom_k] = E_k$, followed by estimating the Pearson correlation between E_k and *T*; **(c)** Reproduction: Chromosomes with greater fitness value have greater chances to be involved in a recombination event. The recombination of chromosomes *x* and *y* constitutes a split at random locus *l*, so that the offspring chromosome is composed of $chrom_x[1{:}l]$ and $chrom_y[(l+1){:}m]$; **(d)** Mutation: Each offspring is subjected to a possible mutation event, where each value in the chromosome can be changed from $0 \rightarrow 1$ or from $1 \rightarrow 0$, respectively. The resulting offspring population is again subjected to a selection process following the same steps. After *g* generations, the last population contains the chromosome with the optimized fitness value. GA, genetic algorithm; WGCNA, weighted gene co-expression network analysis.



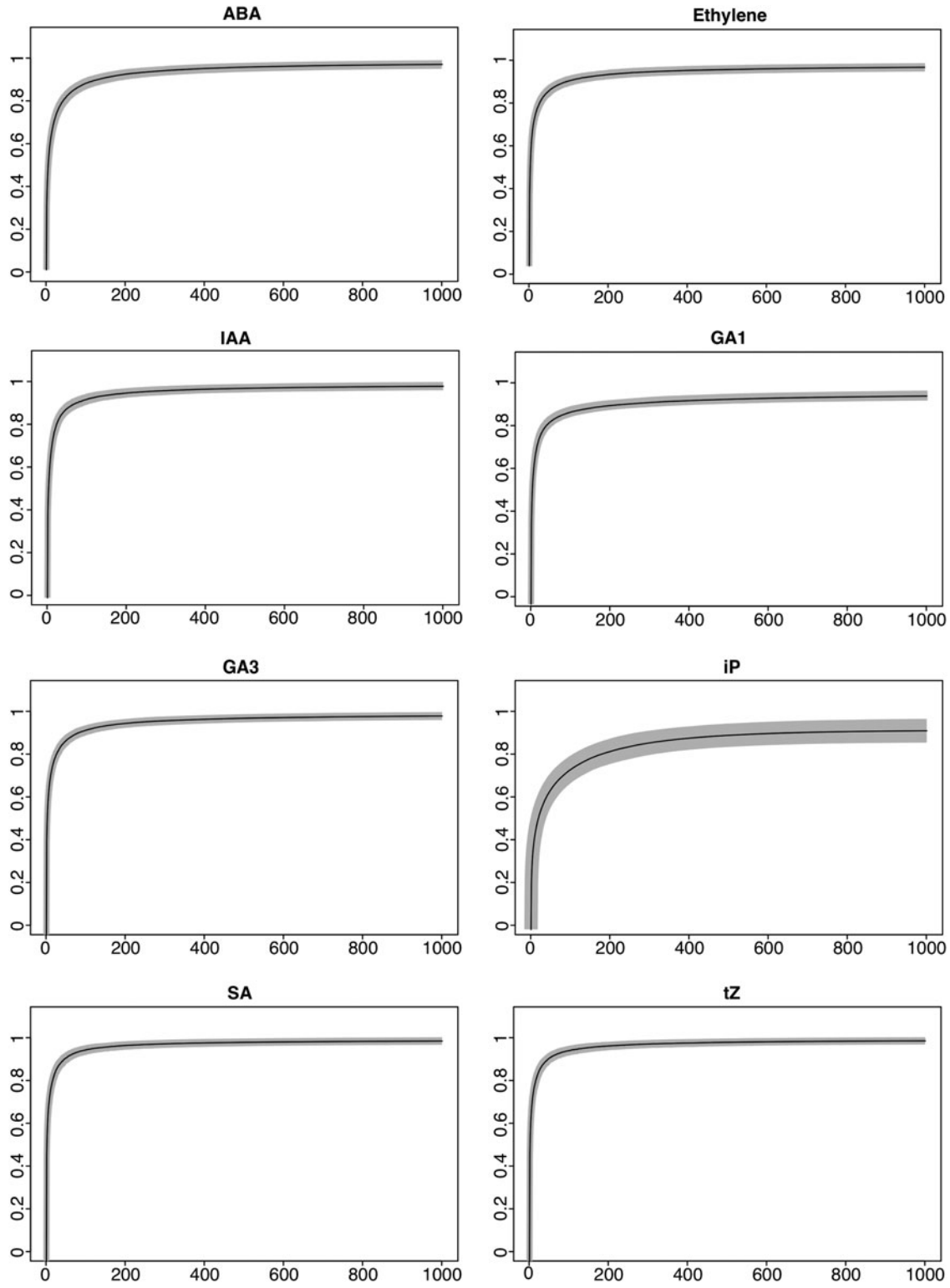


FIG. 3. GA average correlation over 1000 generations and 100 iterations. The GA was performed for 1000 generations and with 100 iterations for all 8 hormones. ABA, ethylene, and IAA were run against CDSs of module darkslateblue, whereas GA_1, GA_3, iP, SA, and tZ were run against CDSs of module turquoise. The average for each generation across iterations was recorded. *X*-axes represent generations, *y*-axes represent correlation coefficients, and gray shaded areas represent standard deviations. ABA, abscisic acid; CDS, coding sequence; GA_1 and GA_3, gibberellins; IAA, indole-3-acetic acid; iP, precursor isopentenyl; SA, salicylic acid; tZ, trans-zeatin.

Table 4. Correlation Overview

Module	*Mean correlation final generation*	*Correlation to ME*	*Strongest correlation to CDS*
Module darkslateblue			
ABA	0.97	0.64	0.91
Ethylene	0.98	0.79	0.9
IAA	0.98	0.77	0.94
Module turquoise			
GA_1	0.94	0.73	0.96
GA_3	0.98	0.79	0.96
iP	0.91	0.36	0.62
SA	0.98	0.94	0.92
tZ	0.98	0.96	0.95

ABA, abscisic acid; GA, genetic algorithm; GA_1 and GA_3, gibberellins; IAA, indole-3-acetic acid; iP, precursor isopentenyl; ME, module eigengenes; SA, salicylic acid; tZ, trans-zeatin.

threshold settings were defined to be relative to the outcome; that is, given 1000 chromosomes with binary values (genes) in the final generation of each iteration: (1) a summation vector of length chromosome with the summed-up value for each gene for all chromosomes of the final generation over all iterations;

$$\text{chromosome}_{\text{sum}} = \sum_{f} \sum_{k} \text{chromosome}_{g^{\text{final}}kf}$$

where f represents iteration, k chromosome, and g generation,

(2) next, duplicate values of the summation vector were removed as a function of:

$$\theta(\text{chromosome}_{\text{sum}}) = \text{unique}(\text{chromosome}_{\text{sum}})$$

(3) finally, the threshold was defined as the average of the unique values:

$$\text{thresh} = \frac{\sigma\theta(\text{chromosome}_{\text{sum}})}{n(\theta(\text{chromosome}_{\text{sum}})}$$

where n detects the length of the vector.

CDSs above the threshold were included into the final subset of the modules' CDSs—the corresponding genes of the modules and the subset of corresponding genes identified for the eight hormones can be viewed in Supplementary Data S1. For ABA, the GA optimized for 213 CDSs, for ethylene 197 CDSs, and for IAA 190 CDSs in contrast to the 2142 CDSs in module darkslateblue. For GA_1, 211 CDSs were identified; for GA_3 284 CDSs; for iP 183 CDSs; for SA 313 CDSs; and for tZ 303 CDSs in contrast to 4437 CDSs in module turquoise.

3.11. Comparative genetic algorithm performance evaluation

To evaluate the performance of the CDS subsets determined by the GA in comparison to other subsets of CDSs of the respective modules, we correlated the first principal component of the final CDS subsets, determined by the threshold settings as described earlier (similar to step b in Fig. 2), to their respective hormones. For ABA, an absolute correlation coefficient of 0.98 was computed: for ethylene 0.98, for IAA 0.98, for GA_1 0.96, for GA_3 0.99, for iP 0.94, for SA 0.99, and, finally, for tZ 0.99 (Fig. 4). Subsequently, empirical p-value analysis was employed, where for 100,000 random subsets of CDSs from modules darkslateblue and turquoise the correlation coefficient of their first principal component to their respective hormones was estimated. The number of CDSs for the random subsets corresponded to the number of CDSs present in the final subset, as determined for each hormone (see Section 3.10). Unequivocally, the subsets determined by the GA recorded higher correlation coefficients than any of the random subsets (Fig. 4).

Next, we pairwise correlated hormone profiles to each CDS expression profile of their respective modules. Then, we chose the CDS profiles with the strongest correlation coefficients and correlated their first principal component to their respective hormones. Again, the number of CDSs with the greatest correlation corresponded to the number of CDSs present in the final subset, as determined for

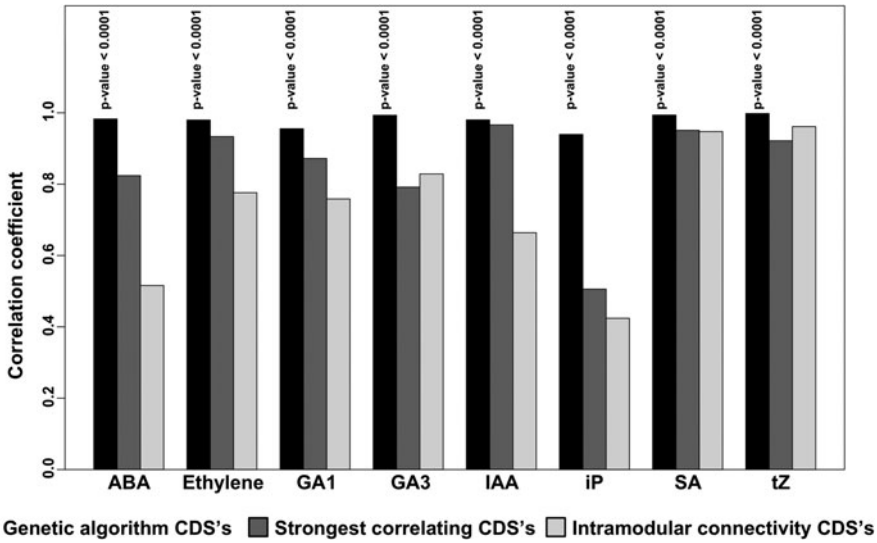


FIG. 4. Comparative performance evaluation. Bar-graph representation of the correlation coefficient of the first principal component of the subsets of CDSs versus their respective hormones. The analysis was performed for the subsets determined by the GA (black bars), by the strongest correlating CDSs (dark grey bars), and intramodular connectivity (light gray). Empirical *p*-value analysis was performed for subsets of CDSs determined by GA. The corresponding values are indicated above the black bars. *X*-axes represent hormones, and *y*-axes represent correlation coefficients.

each hormone. Also, the correlation coefficient for the subset of CDSs determined by the GA was invariably greater than the correlation coefficient for the subsets of CDSs determined by pairwise correlation.

WGCNA includes an approach for determining the most contributing genes within a module. The approach is based on using the weights of the correlation coefficients (edges in a graph) for each node within the module, where nodes with the greatest accumulative weight are considered the most contributing. The value derived from this approach is termed ''intramodular connectivity'' (Dong and Horvath, 2007). We determined the CDSs with the greatest intramodular connectivity values and correlated their first principal component to their respective hormones. The number of CDSs based on the intramodular connectivity corresponded to the number of CDSs present in the final subset for each hormone. The subsets of CDSs determined by the GA recorded greater correlation coefficients than the subsets ascertained by intramodular connectivity (Fig. 4).

3.12. Gene Ontology term enrichment analysis emphasized specificity of genetic algorithm

To provide a biological perspective to the GA outcomes, we performed GO term enrichment analysis for (1) all eight sets of genes generated by the GA, (2) all eight sets of the highest correlating CDSs, (3) and intramodular connectivity. Based on the GA (Table 5), 65 different GO terms were identified, of which 20 were associated with hormones based on module darkslateblue. Out of the 20 identified GO terms, none intersected between hormones ABA, ethylene (Fig. 5a). For hormones based on module turquoise, 49 different GO terms were identified (Table 5), of which 18 (37.5%) intersected between hormones GA_1, GA_3, iP, SA, and tZ (Fig. 5a). For the analysis of the strongest correlating CDSs, 75 significant GO terms were detected (Supplementary Data S2). For hormones based on module darkslateblue, 24 GO terms were identified, of which 1 (4.17%) intersected (Fig. 5b). For hormones based on module turquoise, 59 GO terms were identified, of which 51 (86.44%) intersected (Fig. 5b). For CDSs determined by the intramodular connectivity, 119 significant GO terms were detected (Supplementary Data S2). For hormones based on module darkslateblue, 65 GO terms were identified, of which 61 (93.85%) intersected (Fig. 5c). For hormones based on module turquoise, 57 GO terms were identified, of which 51 (92.98%) intersected (Fig. 5c). The high intersect of GO terms for intramodular connectivity stems from the fact that the identification of significant genes within a module is based on the module itself rather than from its relation to the trait.

Table 5. Significant Gene Ontology Terms Associated with Hormones Based on Genetic Algorithm

GO term	*ABA,* p-*value*	*Ethylene,* p-*value*	*GA_1,* p-*value*	*GA_3,* p-*value*	*IAA,* p-*value*	*iP,* p-*value*	*SA,* p-*value*	*tZ,* p-*value*
Protein localization	**0.00057**	NA	NA	NA	NA	NA	0.067	**0.00034**
Establishment of protein localization	**0.00046**	NA	NA	NA	NA	NA	0.059	**0.00026**
Establishment of localization	**5.00E-04**	0.11	0.63	0.28	0.044	0.86	0.0069	0.018
Localization	**0.00054**	0.11	0.64	0.28	0.046	0.87	0.0074	0.019
Macromolecule localization	**0.0011**	NA	NA	NA	NA	NA	0.097	**0.00075**
Cellular protein localization	**0.0011**	NA	NA	NA	NA	NA	0.033	**0.0019**
Transport	**0.0012**	0.11	0.62	0.27	0.077	0.86	0.0064	0.017
Cellular macromolecule localization	**0.0011**	NA	NA	NA	NA	NA	0.033	**0.0019**
Cellular localization	**0.0022**	NA	NA	0.04	NA	NA	0.055	0.0043
Protein transport	**0.002**	NA	NA	NA	NA	NA	0.051	**0.00019**
Vesicle-mediated transport	**0.0033**	NA	NA	NA	NA	NA	**0.00019**	0.018
Intracellular protein transport	**0.0039**	NA	NA	NA	NA	NA	0.023	**0.0011**
Single-organism cellular process	0.048	**0.00019**	0.57	0.19	0.17	0.034	0.085	0.044
Single-organism catabolic process	NA	**0.00051**	NA	NA	NA	NA	NA	NA
Catabolic process	NA	**0.00057**	NA	NA	0.0084	NA	NA	0.13
Generation of precursor metabolites and energy	NA	**8.00E-04**	NA	NA	NA	NA	NA	NA
Organic substance catabolic process	NA	**0.0021**	NA	NA	0.0069	NA	NA	0.12
Oxoacid metabolic process	0.032	**0.0033**	NA	0.12	NA	0.012	0.3	0.29
Intracellular membrane-bounded organelle	0.084	0.17	**0.00074**	0.065	0.13	0.71	0.27	0.35
Intracellular part	0.024	0.017	**0.00082**	0.0041	0.058	0.031	**7.60E-05**	**0.00089**
Intracellular	0.01	0.025	**3.00E-04**	**0.0038**	0.078	0.007	**3.50E-05**	**0.00043**
Membrane-bounded organelle	0.084	0.17	**0.00074**	0.065	0.13	0.71	0.27	0.35
Cell part	0.012	0.016	**0.00094**	0.01	0.13	0.0084	**7.80E-05**	**0.0016**
Cell	0.012	0.016	**0.00094**	0.01	0.13	0.0084	**7.80E-05**	**0.0016**
Cytoplasm	0.039	0.09	**0.0012**	**0.0024**	0.13	**0.00046**	**5.20E-09**	**1.60E-05**
Intracellular organelle	0.11	0.2	**0.0018**	0.0046	0.24	0.3	**0.00051**	0.011
Organelle	0.11	0.2	**0.0018**	0.0047	0.24	0.3	**0.00051**	0.011
Organelle organization	NA	NA	**0.0043**	NA	NA	NA	NA	0.02
Organelle part	0.058	0.14	0.02	**0.0027**	0.38	0.26	0.03	0.054
Intracellular organelle part	0.058	0.14	0.019	**0.0027**	0.38	0.25	0.03	0.053
Cytoplasmic part	0.12	0.39	0.01	**0.00075**	0.21	0.013	**1.90E-08**	**7.60E-06**
Macromolecular complex	0.068	0.079	0.073	**0.002**	0.058	0.041	**1.30E-05**	**2.30E-05**
Nitrogen compound metabolic process	0.11	0.059	0.34	**0.0017**	0.056	0.0065	0.052	0.0057
Peptide metabolic process	NA	NA	0.092	**0.0039**	0.12	**0.00024**	**3.00E-08**	**0.00084**
Cellular nitrogen compound metabolic process	0.15	0.08	0.31	**0.0018**	0.12	0.017	0.052	0.0096
Structural molecule activity	NA	NA	0.14	**0.003**	NA	0.042	**9.50E-08**	**0.0015**
Cellular macromolecule catabolic process	NA	NA	NA	NA	**0.0013**	NA	NA	NA
Macromolecule catabolic process	NA	NA	NA	NA	**0.0042**	NA	NA	NA
Organonitrogen compound metabolic process	0.17	0.19	0.55	0.017	0.042	**2.30E-05**	**5.50E-07**	**0.00023**
Cellular amide metabolic process	NA	NA	0.096	0.0043	0.12	**0.00026**	**3.70E-08**	**0.00094**

(*continued*)

TABLE 5. (CONTINUED)

GO term	*ABA,* p-*value*	*Ethylene,* p-*value*	*GA_1,* p-*value*	*GA_3,* p-*value*	*IAA,* p-*value*	*iP,* p-*value*	*SA,* p-*value*	*tZ,* p-*value*
Small-molecule metabolic process	0.062	0.03	0.2	0.25	0.23	**0.00044**	0.12	0.11
Peptide biosynthetic process	NA	NA	0.17	0.0086	0.11	**7.00E-04**	**1.90E-08**	**0.00065**
Amide biosynthetic process	NA	NA	0.17	0.0086	0.11	**7.00E-04**	**1.90E-08**	**0.00065**
Translation	NA	NA	0.17	0.0081	0.1	**0.00066**	**1.60E-08**	**6.00E-04**
Cellular amino acid metabolic process	NA	NA	NA	NA	NA	**0.001**	NA	NA
Organonitrogen compound biosynthetic process	0.36	0.57	0.47	0.023	0.05	0.0048	**6.20E-08**	**0.0014**
Ribosome	NA	NA	NA	0.017	NA	0.086	**4.00E-06**	0.0086
Intracellular ribonucleoprotein complex	NA	NA	0.18	0.039	NA	0.14	**2.50E-05**	0.022
Ribonucleoprotein complex	NA	NA	0.18	0.039	NA	0.14	**2.50E-05**	0.022
Intracellular nonmembrane-bounded organelle	0.42	NA	0.34	0.04	NA	0.15	**4.10E-05**	**0.0022**
Nonmembrane-bounded organelle	0.42	NA	0.34	0.04	NA	0.15	**4.10E-05**	**0.0022**
Structural constituent of ribosome	NA	NA	NA	0.019	NA	0.092	**4.90E-06**	0.0096
Biosynthetic process	0.12	0.38	0.73	0.01	0.14	0.035	**0.00076**	0.17
Organic substance biosynthetic process	0.18	0.39	0.65	0.0086	0.23	0.061	**0.0011**	0.16
Cellular biosynthetic process	0.23	0.36	0.71	0.011	0.2	0.082	**0.0014**	0.13
Cellular macromolecule biosynthetic process	0.32	0.47	0.49	0.0065	0.29	0.039	**0.0019**	0.1
Macromolecule biosynthetic process	0.32	0.47	0.49	0.0066	0.29	0.04	**0.0019**	0.1
Carbohydrate derivative biosynthetic process	NA	NA	NA	NA	NA	NA	**0.0034**	NA
Protein complex	0.021	0.056	0.17	0.019	0.093	0.12	0.035	**0.00037**
GTP binding	NA	0.016	NA	NA	NA	NA	0.0073	**0.00039**
Guanyl ribonucleotide binding	NA	0.016	NA	NA	NA	NA	0.0073	**0.00039**
Guanyl nucleotide binding	NA	0.018	NA	NA	NA	NA	0.0084	**0.00047**
Intracellular transport	0.0075	NA	NA	0.029	NA	NA	0.04	**0.0027**
Establishment of localization in cell	0.0075	NA	NA	0.029	NA	NA	0.04	**0.0027**
Organic substance transport	0.016	NA	NA	NA	NA	NA	0.091	**0.0037**

Significant GO terms for respective hormones are highlighted in bold.
GO, Gene Ontology.

3.13. Genetic algorithm identified genes specifically associated to hormones

When inspecting the genes identified by the GA (Supplementary Data S3), a number of genes found are highly associated with hormone regulation, for example, *Prupe.1G349500*, coding for the ABA-responsive GRAM domain-containing protein, functioning downstream of ABI5 (Mauri et al., 2016); *Prupe.2G320300*, encoding a heavy metal transport/detoxification superfamily protein (IAA has been associated with increased heavy metal levels in *Brassica juncea*) (Srivastava et al., 2013) in barley root tips (Zelinova et al., 2015); *Prupe.7G031400*, encoding a member of the auxin carrier family protein (Forestan and Varotto, 2012; Forestan et al., 2012); and *Prupe.3G161500*, encoding a gibberellin-regulated family protein, demonstrated to be involved in plant development (Zhong et al., 2015; Qu et al., 2016; Trapalis et al., 2017). In addition, four cytokinin-associated disease resistance proteins were detected (*Prupe.2G066600*, *Prupe.7G138300*, *Prupe.8G179400*, and *Prupe.2G055700*) (Choi et al., 2010; Grosskinsky et al., 2011; Argueso et al., 2012; Großkinsky et al., 2013).

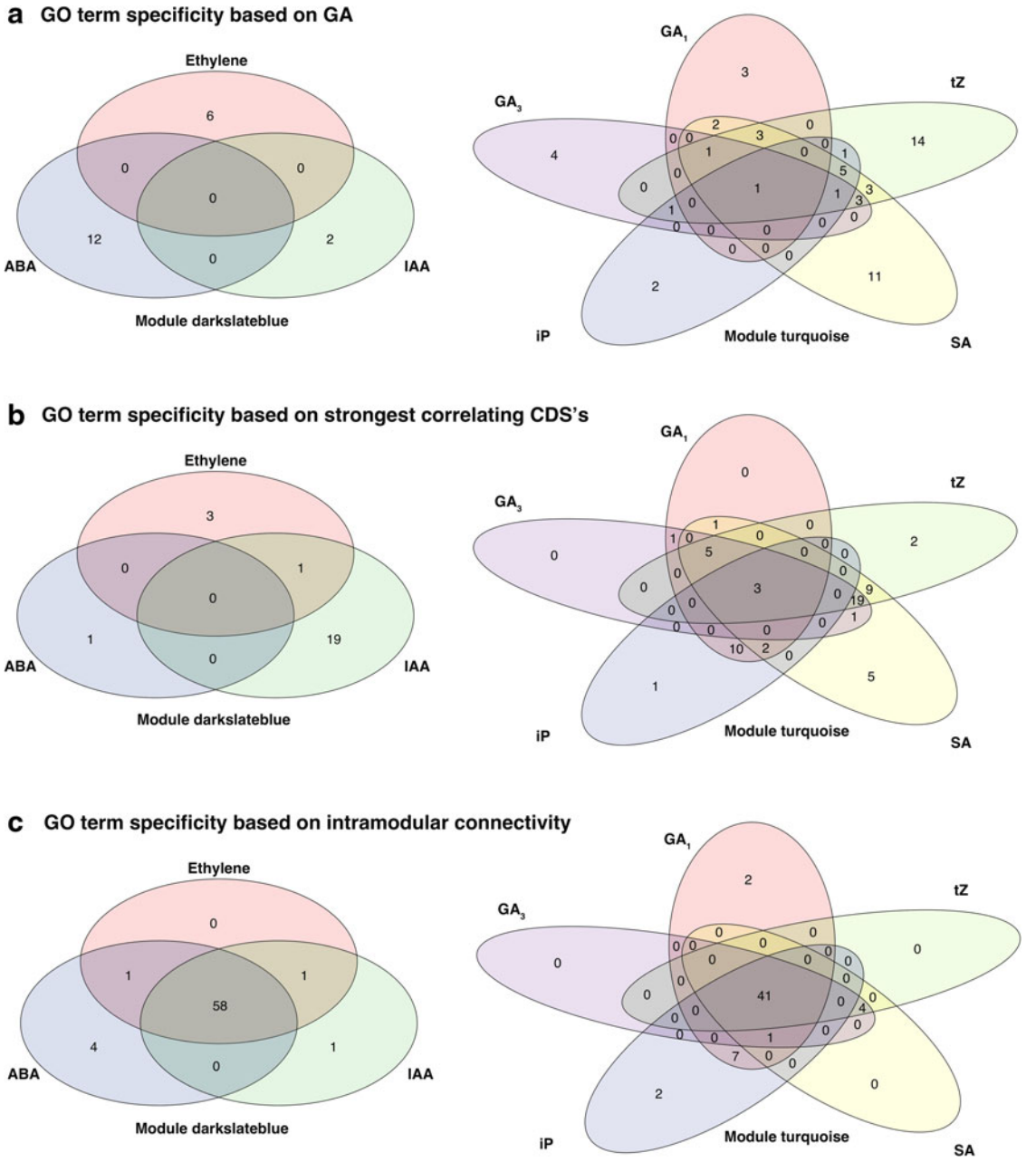


FIG. 5. Intersect of GO terms associated with hormones. Illustrated are three sets of venn diagrams associated with the different methods for the identification of CDSs: **(a)** GO terms determined by the GA; **(b)** GO terms determined by the strongest correlating genes; **(c)** GO terms determined by intramodular connectivity. Venn diagrams of GO terms associated with ABA, ethylene (ETH), and IAA are illustrated on the left (module darkslateblue); venn diagrams of GO terms associated with GA_1, GA_3, iP, SA, and tZ are illustrated on the right (module turquoise). GO, Gene Ontology.

4. DISCUSSION

The identification of genes controlling phenotypic traits is one of the core goals in functional biology. A useful tool to accomplish this task is correlation-based network analysis, where gene expression and trait performance are quantified at different conditions and their co-ordinated behavior is expressed via a correlation coefficient. WGCNA is a useful tool for establishing gene-trait relationships, generating gene modules whose the first principal component is correlated to the trait under investigation. However, often

these gene modules contain thousands of different genes, rendering the gene-trait relationships insufficiently determined to indicate a genuine functional association.

Here, we developed a GA-optimizing gene module to trait relationships, gradually increasing the correlation between the trait and a subset of genes comprising the module. We exemplified the GA on a Japanese plum dataset, where gene expression and content levels of eight different hormones were measured in fruits at different developmental stages in two genetically related cultivars. A comparison between the original correlation coefficient of MEs to hormones (Fig. 1) and the correlation coefficient of the gene subsets within the gene module to hormones (Fig. 3) showed that the GA improved the gene to trait relationships. Further, the comparison between genes detected via the GA and genes proposed by the strongest correlating genes, as well as the WGCNA integrated method intramodular connectivity, revealed that the approach presented in this study consistently outperformed the other two methods. Although the strongest correlating genes method also achieved relatively good results, it showed high fluctuations (Fig. 4).

In parallel, the GA succeeded in significantly reducing the number of genes associated with the respective hormones (Supplementary Data S1). The GA also discriminated the GO terms associated with the eight hormones (Tables 2, 3, 5, Fig. 5 and Supplementary Data S2). Assessment of the set of genes associated with each hormone identified single genes with defined function in hormone regulation (Supplementary Data S3).

5. CONCLUSIONS

We demonstrated that the GA developed in this study is a valuable extension to WGCNA, reducing the number of correlated genes to a number sufficiently small for the assessment of individual genes, thus identifying meaningful candidate genes for subsequent in vivo analyses. We exemplified our study on a fruit hormone and gene expression dataset, but we emphasize that this approach can be used on datasets of any origin (similar to WGCNA itself).

ACKNOWLEDGMENTS

This work was supported by the Will W. Lester Endowment from the University of California.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Data S1
Supplementary Data S2
Supplementary Data S3

REFERENCES

Anders, S., and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.

Anders, S., Pyl, P.T., and Huber, W. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169.

Argueso, C.T., Ferreira, F.J., Epple, P., et al. 2012. Two-component elements mediate interactions between cytokinin and salicylic acid in plant immunity. *PLoS Genet.* 8, e24798.

Choi, J., Huh, S.U., Kojima, M., et al. 2010. The cytokinin-cctivated transcription factor ARR2 promotes plant immunity via TGA3/NPR1-dependent salicylic acid signaling in Arabidopsis. *Dev. Cell.* 19, 284–295.

Dobin, A., Davis, C.A., Schlesinger, F., et al. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Dong, J., and Horvath, S. 2007. Understanding network concepts in modules. *BMC Syst. Biol.* 1, 24.

Du, Z., Zhou, X., Ling, Y., Zhang, Z.H., et al. 2010. agriGO: A GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 38, W64–W70.

Farcuh, M., Li, B.S., Rivero, R.M., et al. 2017. Sugar metabolism reprogramming in a non-climacteric bud mutant of a climacteric plum fruit during development on the tree. *J. Exp. Bot.* 68, 5813–5828.

Farcuh, M., Rivero, R.M., Sadka, A., et al. 2018. Ethylene regulation of sugar metabolism in climacteric and non-climacteric plums. *Postharvest Biol. Tech.* 139, 20–30.

Farcuh, M., Toubiana, D., Sade, N., et al. 2019. Hormone balance in a climacteric plum fruit and its non-climacteric bud mutant during ripening. *Plant Sci.* 280, 51–65.

Forestan, C., Farinati, S., and Varotto, S. 2012. The maize PIN gene family of auxin transporters. *Front. Plant Sci.* 3, 16.

Forestan, C., and Varotto, S. 2012. The role of PIN auxin efflux carriers in polar auxin transport and accumulation and their effect on shaping maize development. *Mol. Plant.* 5, 787–798.

Gondro, C., and Kinghorn, B.P. 2007. A simple genetic algorithm for multiple sequence alignment. *Genet. Mol. Res.* 6, 964–982.

Großkinsky, D., Edelsbrunner, K., Pfeifhofer, H., et al. 2013. Cis- and trans-zeatin differentially modulate plant immunity. *Plant Signal. Behav.* 8, e24798.

Grosskinsky, D.K., Naseem, M., Abdelmohsen, U.R., et al. 2011. Cytokinins mediate resistance against *Pseudomonas syringae* in tobacco through increased antimicrobial phytoalexin synthesis independent of salicylic acid signaling. *Plant Physiol.* 157, 815–830.

Harris, M.A., Deegan, J.I., Lomax, J., et al. 2008. The Gene Ontology project in 2008. *Nucleic Acids Res.* 36, D440–D444.

Kim, H.Y., Saha, P., Farcuh, M., et al. 2015. RNA-Seq analysis of spatiotemporal gene expression patterns during fruit development revealed reference genes for transcript normalization in plums. *Plant Mol. Biol. Rep.* 33, 1634–1649.

Lachmann, A., Torre, D., Keenan, A.B., et al. 2018. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* 9.

Langfelder, P., and Horvath, S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559.

Mauri, N., Fernandez-Marcos, M., Costas, C., et al. 2016. GEM, a member of the GRAM domain family of proteins, is part of the ABA signaling pathway. *Sci. Rep.* 6.

Miller, J.A., Horvath, S., and Geschwind, D.H. 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci U S A.* 107, 12698–12703.

Mitchell, M. 1996. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

Notredame, C., and Higgins, D.G. 1996. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res.* 24, 1515–1524.

Oldham, M.C., Horvath, S., and Geschwind, D.H. 2006. Conservation and evolution of gene colexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci U S A.* 103, 17973–17978.

Qu, J., Kang, S.G., Hah, C., et al. 2016. Molecular and cellular characterization of GA-Stimulated Transcripts GASA4 and GASA6 in *Arabidopsis thaliana*. *Plant Sci.* 246, 1–10.

Srivastava, S., Srivastava, A.K., Suprasanna, P., et al. 2013. Identification and profiling of arsenic stress-induced microRNAs in *Brassica juncea*. *J. Exp. Bot.* 64, 303–315.

Tian, T., Liu, Y., Yan, H.Y., et al. 2017. agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122–W129.

Trapalis, M., Li, S.F., and Parish, R.W. 2017. The Arabidopsis GASA10 gene encodes a cell wall protein strongly expressed in developing anthers and seeds. *Plant Sci.* 260, 71–79.

Vanbatenburg, F.H.D., Gultyaev, A.P., and Pleij, C.W.A. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* 174, 269–280.

Verde, I., Abbott, A.G., Scalabrin, S., et al. 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494.

Verde, I., Jenkins, J., Dondini, L., et al. 2017. The Peach v2.0 release: High-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics.* 18.

Wadapurkar, R.M., and Vyas, R. 2018. Computational analysis of next generation sequencing data and its applications in clinical oncology. *Int. J. Med. Inform.* 11, 75–82.

Wong, K.C., Peng, C.B., Wong, M.H., et al. 2011. Generalizing and learning protein-DNA binding sequence representations by an evolutionary algorithm. *Soft Comput.* 15, 1631–1642.

Zelinova, V., Alemayehu, A., Bocova, B., et al. 2015. Cadmium-induced reactive oxygen species generation, changes in morphogenic responses and activity of some enzymes in barley root tip are regulated by auxin. *Biologia* 70, 356–364.

Zhong, C.M., Xu, H., Ye, S.T., et al. 2015. Gibberellic acid-stimulated arabidopsis6 serves as an integrator of gibberellin, abscisic acid, and glucose signaling during seed germination in Arabidopsis(1 OPEN). *Plant Physiol.* 69, 2288–2303.

Address correspondence to:
Dr. Eduardo Blumwald
Department of Plant Sciences
University of California
1 Shields Avenue
Davis, CA 95616

E-mail: eblumwald@ucdavis.edu