Genome-wide analysis of cis-regulatory element structure and discovery of motif-driven gene co-expression networks in grapevine

Darren Chern Jan Wong¹,*,†, Rodrigo Lopez Gutierrez¹,†, Gregory Alan Gambetta², and Simone Diego Castellarin¹,*

¹Wine Research Centre, University of British Columbia, Vancouver, BC, Canada, and ²Ecophysiologie et Génomique Fonctionnelle de la Vigne, Bordeaux Sciences Agro, INRA, Université de Bordeaux, 33140 Villenave d’Ornon, France

*To whom correspondence should be addressed. Tel. +1 604 827 2667. Email: wongdcj@gmail.com (D.C.J.W.); simone.castellarin@ubc.ca (S.D.C.)

†These authors contributed equally to this study.

Abstract

Coordinated transcriptional and metabolic reprogramming ensures a plant’s continued growth and survival under adverse environmental conditions. Transcription factors (TFs) act to modulate gene expression through complex cis-regulatory element (CRE) interactions. Genome-wide analysis of known plant CREs was performed for all currently predicted protein-coding gene promoters in grapevine (Vitis vinifera L.). Many CREs such as abscisic acid (ABA)-responsive, drought-responsive, auxin-responsive, and evening elements, exhibit bona fide CRE properties such as strong position bias towards the transcription start site (TSS) and over-representation when compared with random promoters. Genes containing these CREs are enriched in a large repertoire of plant biological pathways. Large-scale transcriptome analyses also show that these CREs are highly implicated in grapevine development and stress response. Numerous CRE-driven modules in condition-specific gene co-expression networks (GCNs) were identified and many of these modules were highly enriched for plant biological functions. Several modules corroborate known roles of CREs in drought response, pathogen defense, cell wall metabolism, and fruit ripening, whereas others reveal novel functions in plants. Comparisons with Arabidopsis suggest a general conservation in promoter architecture, gene expression dynamics, and GCN structure across species. Systems analyses of CREs provide insights into the grapevine cis-regulatory code and establish a foundation for future genomic studies in grapevine.

Key words: gene co-expression network, cis-regulatory element, grapevine, promoter, transcriptional regulation

1. Introduction

Transcription factors (TFs) regulate the transcription of target genes both spatially and temporally through the specific binding of cis-regulatory elements (CREs or ‘motifs’) present in their promoters. On average 5% of the protein-coding genes encode TFs in monocotyledon and dicotyledonous plants, which can be divided into super...
families, and further into families, based on their DNA-binding domains.\(^1\) To date, 58 families have been assigned for plants,\(^1\) some of them are grouped in six TF super families. Many of these TFs are essential for plant cell division, development, growth, and responses to the environment.\(^2\) In the last decade, ongoing efforts in curating reported plant CREs\(^3,4\) have been fruitful, providing useful clues into the role of TFs in regulating specific promoter sequences of interest. There have been recent efforts to elucidate the TF-binding specificity in Arabidopsis at a genome-wide scale using \textit{in vitro} techniques coupled to high-throughput platforms such as protein binding microarrays.\(^5\) This strategy provided valuable CRE information for a large number of TFs (63 in total) covering 25 TF families. The most widely studied CREs in plants include the ABRE (abscisic acid (ABA)-responsive element) and DRE (dehydration-responsive element)/CRT (C-repeat) elements which are largely implicated in stress responses,\(^6\) CARG-box,\(^7\) AuxRE,\(^8\) and E2F-binding elements\(^9\) in plant development, and MBS elements implicated in a myriad of development processes and stress responses.\(^10\) Others CREs that belong to lesser known TF families (e.g. Golden2-like, GLK)\(^11\) or that are critical for the binding of well characterized TFs (e.g. AUXIN RESPONSE FACTOR, ARF)\(^5,11\) have been recently characterized as well.

Knowledge pertaining to TF and CRE repertoire has been instrumental in developing regulatory models explaining changes in gene expression patterns across organs, tissues, or in response to stress in Arabidopsis\(^5,2,4-14\) and some crop species.\(^15,16\) Approaches that consider the absence/presence of CREs, the distribution of CRE positions, or their over-representation (ratio of observed/random expectation) have all been used to establish transcriptional regulatory models in plants\(^13,15,17\) and mammals.\(^18\) Many of these studies have highlighted that CREs having a position bias towards the transcription start site (TSS) are more likely to be \textit{bona fide} CREs functioning in important biological roles. Other sophisticated models that consider CRE combinatorial rules\(^12,16\) or its integration with gene co-expression networks (GCNs)\(^15,2,4-12,14\) have also been favored. In GCNs, genes are represented as nodes, and edges connecting any two nodes depict similarity in expression profile across different organs/tissues, developmental stages, and/or stress conditions. GCNs can be further classified into ‘condition-independent’ or ‘condition-specific’ GCNs, representing relationships inferred from multiple experimental conditions or in determined conditions (e.g. abiotic stress, biotic stress, or tissue-specificity), respectively.\(^19\) Lineage-specific GCNs have also been explored.\(^20,21\) Systems-based approaches, such as the integration of CREs, promoter architecture, and GCNs analysis, are promising methodologies to infer gene function,\(^22\) and are complementary to other network inference methods such as associologs—orthologous gene associations—and genome context similarities.\(^23\)

Grapes (\textit{Vitis vinifera} L.) are amongst the world’s major fruit crops and are highly valued for fresh fruit consumption and winemaking. The grapevine genome\(^24\) provides a strong foundation for elucidating the molecular genetics involved in grapevine physiology and response to environment. Specifically, seminal studies involving large-scale gene expression profiling have revealed important mechanisms and biomarkers critical for grapevine development across organs/tissues\(^25,26\) and berry phenotypic plasticity (e.g. genotype \(\times\) environment interactions).\(^27\) The availability of the grapevine genome and the rapid accumulation of publicly available grapevine whole-genome microarray and RNA-seq datasets provide the raw material necessary for GCN and CRE analyses. To date, several studies have utilized GCN and CRE analyses in revealing biologically relevant target genes of grapevine TFs\(^28-30\) and/or providing clues into the regulatory relationships of TFs in key metabolic and ripening pathways.\(^31,32\) However, to date no study has characterized grapevine promoter architecture and identified putative CRE regulatory modules at a genome-wide scale.

In this study, the distribution and organization of promoter CREs and their properties were analyzed for all currently predicted protein-coding genes in grapevine. Further insights into the biological role(s) of genes containing these CREs were achieved through analysis of enriched plant biological pathways. Meta-analysis of next-generation sequencing (RNA-seq) datasets and the grapevine tissue developmental atlas revealed the importance of many CREs in stress and developmental regulation. A \textit{bottom-up} strategy was adopted to identify regulatory modules potentially driven by CREs in both stress- and development-specific GCNs. We apply a robust approach by integrating different layers of information—promoter architecture, gene expression, and GCNs—to identify the cis-regulatory code during environment stress, growth, and/or development in grapevine. The insights of this study will allow the prioritization of CREs and gene candidates for functional characterization in grapevines.

2. Materials and methods

2.1. Discovery of known plant CREs in grapevine promoters

Candidate plant CREs, were compiled from two plant promoter databases, namely AGRIS\(^3\) and PLACE,\(^1\) and from protein-binding microarray inferred CREs of 25 Arabidopsis TF families.\(^5\) A total of 559 non-redundant CREs were retained for subsequent analysis. The promoters of the grapevine genes were downloaded from EnsemblPlants using Biomart (http://plants.ensembl.org/index.html; 1 March 2016, date last accessed). Sequences between -1 to -1,000 bp upstream of each gene TSS containing < 10% N sequences were deemed as candidate promoter sequence of each gene. All 559 non-redundant CREs were scanned for exact pattern match (no mismatch) across 29,839 promoters on both + and − strands, and indexed for the number of occurrence, strand, and position information.

2.2. Positional bias and fold-change enrichment analysis of CREs in grapevine promoters

Each CRE was evaluated for positional bias using two popular approaches; the clustering factor (\(CF\))\(^22\) and a recently proposed index, the \(Z\)-score.\(^13\) Another approach based on the fold-change/enrichment between observed/expected frequencies with resampling was considered.\(^15\) For the \(CF\) calculation, the number of occurrences of each CRE in each bin (50 bases, 20 bins total) was calculated. Based on the abundance of the CRE in the 20 bins, the mean (\(x\)) and standard deviation (\(s\)) were calculated. A second mean (\(x'\)) and a second standard deviation (\(s'\)) were calculated based only on bin values that were < 2\(x\) above \(x\). Then the \(CF\) is calculated with the equation: \(CF = (x_{\text{max}} - x')/s'\). The \(Z\)-score,\(^13\) for a given CRE was calculated considering every gene in the mutual rank (MR) network(s). The \(Z\)-score for each CRE is calculated with the equation: \(Z\)-score = \((L+2+p)/(L-(1+1)^2-1)/sqrt(n)\), where \(L\) corresponds to the length of the promoter, \(l\) corresponds to the length of the motif, \(n\) corresponds the number of instances that the CRE is found in the promoters of the genes, and \(p\) corresponds to the mean position from all the instances the CRE was found. For the fold-change/enrichment (FE) calculation, the observed frequency is calculated as in the \(Z\)-score, based on the number of instances that the CRE was found in the promoters, and the expected frequency was calculated by
performing 1,000 replicates of random sampling with equal sizes (number of genes containing the selected CRE).

2.3. Compilation of transcriptome datasets, differential expression (DE), and comparative gene expression analysis

Two different compilations of databases were made: a stress-related compendia, which was mainly derived from next-generation sequencing (RNA-seq) datasets publicly available from NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra; last accessed 01/03/2016), and a grapevine tissue developmental series (atlas), based on the NimbleGen Grape Whole-genome microarray platform. For RNA-seq datasets, quality control of raw single- and pair-end FASTQ reads were performed with Trimmomatic v0.36, ensuring all reads maintained a minimum Q score of 20 and a length of 40. Bowtie2 v2.2.7 was used with default parameters for read alignments using the 12x grapevine reference genome PN40024. Read summarization was performed with htsq-count v0.6.1 using the 12x V1 annotation with default settings. DESeq2 was used for the DE analysis of the stress (RNA-seq) datasets, comparing the experiments’ treatments against controls. For the grapevine atlas dataset, DE analysis was performed using limma comparing the development progression from young to mature of each organ, except flower-, seedling-, and root-specific tissues. A threshold of false discovery rate (FDR) < 0.05 and absolute log2FC > 1 represents significant DE genes identified through DESeq2 and limma in each comparison. Estimation of RNA-seq transcript abundance was performed with DESeq2 by applying variance-stabilizing transformation (VST) while pre-normalized probe intensities (using Robust Multi-Array Average), were gathered from. All predicted grapevine-Arabidopsis orthologues were downloaded from EnsemblPlants (http://plants.ensembl.org/index.html; 1 March 2016, date last accessed). Expression analysis of Arabidopsis genes was conducted using the ArGenExpress compendium for the ‘stress’ and ‘pathogen’ series via the Bio-Analytic Resource for Plant biology database (http://bar.utoronto.ca; 1 December 2016, date last accessed) expression browser tool.

2.4. Construction of GCNs

Prior to network construction, genes that are not DE in at least one comparison in each compendium were filtered. A final non-redundant list of DE genes in the stress-responsive and atlas compendia were obtained and used for GCN construction. Principal component analysis was performed on the RNA-seq VST dataset using R (http://www.r-project.org). Pearson’s correlation coefficient (PCC) values were used as a measurement of expression similarity between gene pairs. Ranks were assigned to gene pairs based on the PCC values of their genes. The formula for geometric mean of the ranks, MR(AB) = (Rank(A-B) × Rank(B-A)), was used to calculate the MR value for each pair-wise genes. Rank(A-B) corresponds to the rank assigned to gene B given the list of co-expression genes from gene A, while vice versa for Rank(B-A). For each individual network (stress and atlas), the significance of MR values was estimated based on 1,000 permutations as per. 21

2.5. Network-based CRE subnetwork selection and functional enrichment analysis

For each gene and all immediate connected genes (nodes) forming a group, enrichment P-values for CREs in their promoters were based on hypergeometric distribution. From each GCN with significant MR values, genes having a given CRE enriched at P < 0.01 were then selected as ‘seed’ or ‘guide’ genes to generate the relevant CRE network. For the construction of Arabidopsis CRE network (P < 0.01), the Arabidopsis graphical Gaussian model gene network with a partial correlation threshold > 0.05 was used. To further identify densely connected regions (subnetworks) within each CRE network, the Girvan–Newman fast greedy algorithm was applied. Functional enrichment of MapMan ontology terms for grapevine and Arabidopsis was performed in R and terms with an FDR < 0.05 were considered significantly enriched. In each CRE subnetwork, genes within selected ‘seed’ or ‘guide’ gene(s) node vicinity of up to two steps (k = 2) were extracted for further analyses. All visualization of various CRE subnetworks was performed with Cytoscape v3.4.

3. Results and discussion

3.1. Distribution of CREs in grapevine promoters

An evaluation of base composition is often the first step in characterizing promoter sequences of an organism. In this study, analysis of promoter sequences (1,000 bp upstream of the TSS) revealed an AT-rich ratio, where A:C:G:T composition are 0.33:0.16:0.16:0.34, respectively. Promoter sequences of Arabidopsis, soybean, and rice are also AT-rich which is a strong indicator of bona fide plant promoter regions. Therefore, promoter sequences extracted here likely reflect actual promoter regions and not random genomic regions. Rigorous statistical and DNA-binding specificity evaluation have shown that CREs between 6- and 8-mer provide an optimal threshold for capturing most biologically relevant CREs. Indeed, the occurrence of CREs, shorter than 6-mer or larger than 8-mer, in the promoters of the grapevine genome (Supplementary File 1) might result in a large or small number of hits, respectively, complicating the statistical evaluation for the identification of biologically relevant CREs (Fig. 1, Supplementary Table S1). Focussing on promoter sequences containing CREs between 6- and 8-mer, three separate approaches were considered to identify biologically relevant CREs, fulfilling a positional bias towards the TSS (using CF and Z-score) and/or having high fold enrichment when compared with random occurrences in shuffled promoter sequences (using FE). CREs were seen to score differently according to each of the three approaches (Supplementary Table S1).

Several examples 6-mer CREs that are canonical TF binding sites for auxin responsive factor [AuxRE5,11,42; TGTCTC and TGTCGG], AP2/ERF superfamily (DRE43; RCCGAC and GCCGAC; and GCC-box43; GCCGCGC), bZIP44 (G-box: CACGTG; and C-box: GACGTC), bHLH45,46 (G-box), and CAMTA (CGCGBOXAT7; VCGGCG) TFs showed strong position bias towards the TSS and/or were strongly over-represented. The AuxRE TGTCCT CRE had a CF (highest) and Z-score score of 5.8, was present in 9,111 promoters (n = 11,621 occurrences) with a peak position between 0 and 50 bp from the TSS, and had a FE value showing 3.3-fold more occurrences observed compared with random promoters (Fig. 1, Supplementary Table S1). A novel AuxRE, TGTCCG, which was recently shown to be the preferred binding site for ARFs compared with TGTCCT5,11 has a slightly different distribution profile. TGTCCG is present in fewer (4,411) promoters, with a different peak position (50–100 bp upstream of TSS), lower CF and Z-score, but with higher FE values (6.8-fold more occurrences than in random promoters). Commonly associated bZIP
and bHLH binding site motif, G-box (CACGTG), was present in 6,200 promoters (n = 7,956 occurrences) with a peak position between 50 and 100 bp from the TSS, and CF, Z-score, and FE, of 2.9, 6.1, and 4.8 (Supplementary Table S1). Meanwhile, the GCCGAC CRE, a component of DRE, was among the CREs with the highest FE (12.9-fold more occurrences than in random promoters), was present in 2,310 promoters (n = 2,529 occurrences), with the same peak position as G-box, with CF and Z-score of 2.9 and 3.9, respectively.

Considering 8-mer CREs, several CREs potentially targeted by E2F TF including E2F1OSPCNA (E2F1OSPCNA) and TTTTCGCG (E2FANTRNR), present in 347 (n = 360 occurrences) and 410 (n = 418 occurrences) promoters were ranked highly considering CF, Z-scores, and FE criterions. In addition, several longer instances of G-box (e.g. CACGTGCG and CCACGTGG), and other instances of bZIP TFs (e.g. TACGTGTC) were also highly ranked. Among the top-scoring 8-mer CREs for CF is the AGATATT sequence, a circadian clock associated CRE, recognized by Arabidopsis MYB-related TF, CCA1, and RVE1. There were 4,120 promoters containing AGATATT (n = 4,567 occurrences) with a peak position between 0 and 50 bp from the TSS, and a FE value of 7.2 (Supplementary Table S1).

Top-scoring CREs in grapevine promoters, such as the G-box, CGCGBOXAT, DRE, GCC-box, E2F1OSPCNA, and AGATATT, share similar promoter distribution profiles as in the promoters of other species; having a peak position within 200 bp upstream of TSS, a high Z-score (>3) and/or CF, or are highly over-represented compared with random promoters reaffirming some extent of conservation in plant promoter architecture and CRE distribution. Several exceptions exist where well-characterized plant CREs, such as CArG box (CC[A/T] 6GG) and A-box (TACGTA) recognized by MADS TFs and bZIP TFs, respectively (Fig. 1, Supplementary Table S1), do not possess authentic CRE properties. This observation does not discount these CREs’ involvement in the regulation of plant biological functions, showing the need for additional scanning approaches or other statistical models in order to capture the full spectrum of functional CREs in grapevine.

3.2. Enriched biological pathways support the relevance of CREs in grapevine promoters

To ascertain the biological roles of genes whose promoters contain these CREs, functional enrichment analysis has been widely adopted. In this study, the MapMan ontology, tailored specifically for plants, was adopted for the identification of enriched biological pathways due to its improve gene function prediction performance compared with gene ontology. The evaluation of all compiled CREs was performed and summarized in Supplementary Table S2. Emphasis is given to those which scored highly across the CF, Z-score, or FE (Fig. 2).

The AuxRE, TGTCTC, was enriched (FDR < 0.05) in protein targeting, regulation of transcription, and secondary metabolism with 64,724, and 251 genes, respectively (Supplementary Table S2). Conversely, TGTCCG-containing promoters are highly enriched in hormone metabolism (168 genes, FDR < 5.5E-04) pathways, specifically auxin (49 genes) and ethylene (38 genes) biosynthesis and regulation. Cell wall metabolism (105 genes) and post-translational modification (160 genes) pathways were also significantly enriched. Enrichment of hormone homeostasis and post-translational modification pathways is not surprising as these mechanisms regulate plant ARF function and ARFs are known to facilitate feedback loops within the auxin signalling pathway. Although both TGTCTC’s and TGTCCG’s functional enrichment profiles suggest that various pathways are potentially targeted by ARFs in grapevine, TGTCCG-
containing promoters show higher enrichment profiles coinciding with a higher affinity presence of ARFs. Nonetheless, treatment of the auxin analogue (naphthaleneacetic acid) in grape berries at the onset of ripening and harvest significantly modulated >1,000 genes at each stage, many of which were auxin- and ethylene-related pathway genes, consistent with the enrichment profiles of TGTCGG-containing genes reported here. As no grapevine ARFs have been functionally characterized, the specificity of each member (19 in total) in delineating real targets from the vast number of AuxRE-containing promoters, and the exact mode of their regulation (activation/repression), remains to be determined.

The CACGTG (G-box) containing promoters were enriched (FDR < 0.05) in a plethora of pathways including regulation of transcription (552 genes), abiotic and biotic stress (230 genes), development (185 genes), flavonoid metabolism (85 genes), sugar metabolism (43 genes), and many more (Supplementary Table S2). This observation is consistent with the role of plant bZIP and bHLH TFs; especially a subset of those who regulate a large repertoire of plant development processes and stress responses (e.g. bZIP-HY5 and bHLH-PFs). In grapevine, recent studies demonstrate that bZIP/bHLH TFs have a role in regulating fruit ripening (VviABF2) and the flavonoid pathway (VvibZIPC22, VviHY5, and VviMYC1). The high enrichment of CACGTG-containing genes (122 genes, FDR < 2.1E-05) in biotic stress pathways is of interest as limited information is known on the roles of bZIP and bHLH TFs in plant immunity regulatory networks.

Promoters of genes containing AGATATTT, a circadian clock-associated CRE, were highly enriched in secondary metabolism (197 genes, FDR < 4.6E-06; especially flavonoids and isoprenoids), hormone metabolism (159 genes, FDR < 1.1E-04), and abiotic stress.

**Figure 2.** Enriched functional categories of genes containing selected 6-mer and 8-mer CREs in their promoter sequences. The opacity of the dot corresponds to –log10 FDR functional category enrichment values. The size of the dots corresponds to the number of genes (log10 scaled) whose promoters contain the selected CRE. Commonly associated names for CREs are indicated in brackets.
Numerous CREs are implicated in grapevine stress-responsive and development-dynamic transcription

To determine the contribution of CREs in determining stress-responsive and development-dynamic transcription, meta-analysis of publicly available next-generation sequencing (RNA-seq) and the grapevine tissue developmental atlas was first performed (Supplementary Fig. S1). The survey for stress-related RNA-seq datasets yielded a total of 87 conditions (260 non-averaged) from 10 experiments that considered various abiotic (e.g. drought and shading) and biotic stress (e.g. Botrytis cinerea, Erysiphe necator, and Plasmopara viticola) (Supplementary Table S3). All these conditions also correspond to three different organs: leaf, berry, and flower. Given the incompleteness of sequenced mRNA across tissues and associated developmental stages, the grapevine expression atlas was considered.

3.3. Numerous CREs are implicated in grapevine stress-responsive and development-dynamic transcription

To determine the contribution of CREs in determining stress-responsive and development-dynamic transcription, meta-analysis of publicly available next-generation sequencing (RNA-seq) and the grapevine tissue developmental atlas was first performed (Supplementary Fig. S1). The survey for stress-related RNA-seq datasets yielded a total of 87 conditions (260 non-averaged) from 10 experiments that considered various abiotic (e.g. drought and shading) and biotic stress (e.g. Botrytis cinerea, Erysiphe necator, and Plasmopara viticola) (Supplementary Table S3). All these conditions also correspond to three different organs: leaf, berry, and flower. Given the incompleteness of sequenced mRNA across tissues and associated developmental stages, the grapevine expression atlas was considered.

Based on the NimbleGen microarray platform capable of profiling > 95% of the predicted grapevine transcriptome. The grapevine atlas represents the most comprehensive (54 tissues) compendia of tissue x developmental gene expression series to date.

In the stress-related compendia, 17,623 genes (59% of transcriptome) were collectively modulated (log2FC > 1, FDR < 0.05) across all the comparisons (Supplementary Table S4 and Supplementary Fig. S1). Abiotic stress treatments (drought and shading) modulated more than 5,000 genes in leaves and flowers, while biotic stress treatments (B. cinerea, E. necator, P. viticola, and Tetranychus urticae) modulated more than 4,000 genes collectively in berries and leaves. Previous work by Fasoli et al.22 revealed the grapevine atlas transcriptome relationships to be highly clustered according to their developmental indices, but did not identify the genes that are developmentally modulated in each tissue/organ. Meanwhile, reanalysis of grapevine atlas was performed to identify genes that are modulated during development and maturity progression. As expected, more than 85% (25,947 genes) of the predicted transcriptome is significantly modulated (log2FC > 1, FDR < 0.05) across the whole atlas (excluding carpel, stamen, petal, pollen, seedling, and root). In organs such as berry (flesh, pericarp, and seed), tendril, stem, and buds more than 10,000 genes were modulated as development progresses while berry skin, rachis, leaf, and inflorescence had lesser modulated genes.

Several studies have shown that filtering for highly modulated genes, often within top 100 most modulated genes15 or having large (> 6-fold) fold changes, may be better suited for identifying major CREs driving expression under a particular condition. We empirically determined that the top 300 most up- and down-regulated genes would provide a reasonable threshold closely satisfying the large fold change requirements (Fig. 3A; Supplementary Fig. S2). These genes typically have a median fold change greater than 4- and 8-fold in the stress and development datasets, respectively. In addition, 10,847 and 8,224 unique genes were within top 300 up- and down-regulated genes in at least one stress and developmental condition, respectively, with an overlap of 5,601 genes that were within the 300 up- and down-regulated genes in both stress-responsive and developmental-dynamic condition.

For each 6-mer CRE, the proportion of the genes (containing the respective CRE) that were highly responsive to either stress-responsive and/or development-dynamic was analysed in detail. Considering all 6-mer CREs, their proportions had a median distribution of 46% (Fig. 3B); more specifically, a median distribution of
37% was considered only stress-responsive, 28% was only development-dynamic, while 19% was considered both stress-responsive and development-dynamic (Fig. 3C). Similar proportions were also observed for 8-mer CREs (Fig. 3B and C), albeit some outlier CREs show that genes containing the latter CREs were neither highly stress-responsive nor development-dynamic. Therefore, this suggests that the distribution of 6-mer to 8-mer CREs in gene promoters likely contributes to approximately half of the DE expression (e.g. MR) can be used to infer co-expression relationships, albeit shown that direct correlations (e.g. PCC) and rank-based co-expression are generally more robust and is the most predictable performance compared with interpreting ‘global’ CRE-driven modules, 293 and 314 for stress and development, respectively, demonstrating that CRE-driven submodules have higher predictive performance compared with interpreting ‘global’ CRE-driven modules as a whole. The biological relevance for some of these condition-specific CRE-driven submodules will be discussed in greater detail.

### 3.4. CRE-driven modules are biologically relevant in condition-specific GCNs

Two GCNs (one stress- and one development-specific) were constructed with gene expression matrices using the MR approach. Co-expression studies in many plants have shown that direct correlations (e.g. PCC) and rank-based co-expression (e.g. MR) can be used to infer co-expression relationships, albeit rank-based co-expression is generally more robust and is the most popular approach among crop GCNs. In addition, establishment of statistically significant reciprocal ranks, MR in this case, may further improve the robustness of the resulting GCN by retaining strong and significant co-expression. In the two datasets, we empirically chose a MR cut-off of 35 based on 1,000 permutations of each expression matrices, which corresponds to a statistical confidence of $P < 0.003$ and 0.005 for the surviving MR in the stress- and development-specific GCN, respectively. These network criteria are highly comparable to the parameters applied in other studies that define strong and significant co-expression. For example, GCNs constructed with MR < 0.5, statistical significance of reciprocal ranks with $P < 0.01$, or by retaining top 100 strongest connections. The final stress-specific GCN results in 10,800 nodes and 134,100 edges, while the final development-specific GCN consists of 8,224 nodes, 113,790 edges (Supplementary Tables S5 and S6). The number of connections for each node in the stress-specific GCN network ranges between 1 and 84, with a median of 21 connections, while the development-specific GCN ranged between 2 and 93, with a median of 24 connections.

Next, we adopted a recently proposed ‘bottom-up’ approach by Ma and co-workers to identify gene expression modules potentially driven by CREs of co-expressed genes. Using this approach, 245 and 289 CRE-driven modules (Supplementary Tables S7 and S8) in stress- and development-specific GCNs, respectively, were enriched (FDR < 0.05) with at least one MapMan BIN category (Supplementary Table S2). Although modular structures self-manifested in most CRE-driven modules, we applied community clustering to ensure robust final submodules. In total, 880 and 1,095 CRE-driven submodules were enriched (FDR < 0.05) in the stress and development-specific GCNs, respectively, with at least one MapMan BIN category (Supplementary Table S2). Interestingly, these enriched submodules corresponded to more ‘global’ CRE-driven modules, 293 and 314 for stress and development, respectively, demonstrating that CRE-driven submodules have higher predictive performance compared with interpreting ‘global’ CRE-driven modules as a whole. The biological relevance for some of these condition-specific CRE-driven submodules will be discussed in greater detail.

### Table 1. MapMan enrichment of selected G-box and GCC-core submodules in the stress dataset

<table>
<thead>
<tr>
<th>CRE Module</th>
<th>CRE</th>
<th>x</th>
<th>n</th>
<th>k</th>
<th>P value</th>
<th>FDR</th>
<th>MapMan term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCC-core (G-box)</td>
<td>I</td>
<td>7</td>
<td>189</td>
<td>132</td>
<td>2.19E-05</td>
<td>1.48E-02</td>
<td>Hormone metabolism, abscisic acid</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>5</td>
<td>55</td>
<td>74</td>
<td>2.71E-07</td>
<td>3.68E-04</td>
<td>Cell wall, pectin * esterases</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>9</td>
<td>67</td>
<td>549</td>
<td>3.75E-06</td>
<td>5.09E-03</td>
<td>Cell wall</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>5</td>
<td>57</td>
<td>181</td>
<td>2.55E-05</td>
<td>1.16E-02</td>
<td>Secondary metabolism, phenylpropanoids</td>
</tr>
<tr>
<td></td>
<td>VI</td>
<td>7</td>
<td>163</td>
<td>145</td>
<td>1.56E-05</td>
<td>1.35E-02</td>
<td>Secondary metabolism, flavonoids</td>
</tr>
<tr>
<td></td>
<td>VII</td>
<td>3</td>
<td>7</td>
<td>389</td>
<td>7.61E-05</td>
<td>3.45E-02</td>
<td>Secondary metabolism, flavonoids</td>
</tr>
<tr>
<td></td>
<td>VIII</td>
<td>3</td>
<td>5</td>
<td>498</td>
<td>4.66E-05</td>
<td>3.16E-02</td>
<td>Stress, abiotic</td>
</tr>
<tr>
<td></td>
<td>IX</td>
<td>11</td>
<td>25</td>
<td>288</td>
<td>2.55E-16</td>
<td>6.92E-14</td>
<td>Lipid metabolism, lipid degradation</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>8</td>
<td>25</td>
<td>1,111</td>
<td>2.28E-06</td>
<td>3.09E-04</td>
<td>Stress, abiotic</td>
</tr>
<tr>
<td></td>
<td>XII</td>
<td>3</td>
<td>143</td>
<td>8</td>
<td>3.29E-13</td>
<td>4.47E-10</td>
<td>Lipid metabolism, TAG synthesis</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>6</td>
<td>143</td>
<td>18</td>
<td>8.54E-05</td>
<td>3.87E-02</td>
<td>Development, late embryogenesis abundant</td>
</tr>
<tr>
<td>GCC-core</td>
<td>I</td>
<td>6</td>
<td>101</td>
<td>145</td>
<td>1.04E-05</td>
<td>5.19E-03</td>
<td>Hormone metabolism, ethylene</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>5</td>
<td>94</td>
<td>61</td>
<td>1.50E-06</td>
<td>2.15E-03</td>
<td>RNA, regulation of transcription, WRKY</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>33</td>
<td>71</td>
<td>304</td>
<td>6.49E-47</td>
<td>3.10E-44</td>
<td>Secondary metabolism, flavonoids</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>3</td>
<td>71</td>
<td>174</td>
<td>3.77E-10</td>
<td>7.71E-08</td>
<td>Secondary metabolism, phenylpropanoids</td>
</tr>
<tr>
<td></td>
<td>VI</td>
<td>8</td>
<td>27</td>
<td>1,016</td>
<td>2.25E-06</td>
<td>1.61E-03</td>
<td>Signalling, receptor kinases</td>
</tr>
</tbody>
</table>

x, overlap size; n, query size; and k, term size represent the contingency tables for MapMan enrichment calculations.
3.5. CACGTG enriched co-expression network is involved in ABA metabolism and light-related functions

The G-box (CACGTG) was found to be highly enriched for stress-responsive genes having scored a CF, Z-score, and FE of 2.7, 3.7, and 4.5, respectively. To investigate whether the enrichment observed has a biological relevance, a gene network was constructed for the stress dataset. A total of 842 genes composed a network comprising a total of 20 community clusters. Some of these clusters were significantly enriched with different MapMan BIN categories (Table 1). Focus is given to the core modules with bZIP and bHLH TFs since previous research has shown that the G-box CRE has been associated with the binding of these TFs.44,46 The presence of these TFs within the modules of the CACGTG network would suggest that the TFs contribute to the regulation of genes and processes pertaining to the selected module.

Module I (Fig. 4A and B) had a MapMan ontology enrichment pertaining to ABA metabolism. In this module six bZIPS and one
bHLH were identified, of which one bZIP encodes the grapevine ABSCISIC ACID RESPONSE ELEMENT-BINDING FACTOR2 (VviABF2, VIT_18s0001g10450) involved in regulating berry ripening in an ABA-dependent manner. However, direct implication of VviABF2 in regulating stress response networks remains to be established. In Arabidopsis, previous research has shown that this homologous bZIP is involved in ABA mediated stress responses including drought. In the VviABF2 subnetwork (k = 2) of Module 1 (Fig. 4A), members of theABA signaling pathway were immediately connected with VviABF2 including a homolog (VviABI1, VIT_11s0016g03180) of an Arabidopsis protein phosphatase 2C, ABA insensitive 1, an essential component of the ABA signaling network, and two homeodomain-leucine zipper TFs (VviHB7, VIT_15s0048g02870 and VviHB12, VIT_02s0025g02590), involved in the signal transduction pathways mediating growth in response to drought. In addition, VviABF2 might be involved indirectly (k = 2) in the regulation of a 9-cis-epoxycarotenoid dioxygenase 2 (VviNCED3, VIT_19s0009g00550), encoding the putative rate-limiting step of ABA biosynthesis, and a protein phosphatase 2C gene (VviPP2C9, VIT_06s0004g05460) involved in ABA.
signalling. To determine the involvement of this subnetwork in grapevine stress responses, meta-analysis of stress-related RNA-seq datasets described above was performed. Water deficit resulted in higher gene expression and greater up-regulation of module genes especially in leaves compared with berries (Fig. 4C). This observation is in agreement with the role of ABA metabolism and signaling in response to drought, while highlighting differences in ABA metabolism and signaling pathways that arise between organs. Another bZIP gene in this module, encodes a grapevine HY5 HOMOLOGUE (VviHY5, VIT_05s0022g01090). Our recent work proposed that VviHY5 may fulfill both co-operative and partially redundant roles with VviHY5 during both grapevine photomorphogenic growth, berry development, and under stress. Supporting this, many genes involved in light- and chloroplast-related functions were densely connected the VviHY5 subnetwork (k = 2) of Module I (Fig. 4B). This includes electron transport pathway genes (e.g. NDHT, VIT_18s0117g00260 and PGR5, VIT_05s0049g01760), chlorophyll breakdown (e.g. ACD1/PAO, VIT_06s0004g00610), chloroplast-targeting (TICT5, VIT_04s0008g07020 and ANKYRN REPEAT PROTEIN, VIT_04s0008g03840), DNA-damage repair (e.g. PHOTOLYASE 1/PHR1/UVR2, VIT_02s0241g00040 and CRY3, VIT_04s0008g02670), and light signaling (SUSPENSION OF PHYA 3/SPA3, VIT_12s0028g03570). In addition, genes involved in the glycosylation of flavonoids—essential for enhancing their solubility in water and accumulation in plant cells—namely VviGT5 (VIT_11s0052g01630) and VviGT6 (VIT_11s0052g01600), were connected with VviHY5 within one and two steps, respectively. An Arabidopsis MYB-like domain TF homolog (VviMYB5, VIT_01s0026g01050), involved in the regulation of anthocyanin biosynthesis, was also co-regulated with VviHY5 (k = 2). These observations reaffirm the involvement of VviHY5 submodule genes in light-regulated transcriptional and secondary metabolism networks in a stress-related context. The strong presence of the G-box CRE may imply a co-operative role of other bZIP TFs in its regulation, such as HY5. This has been shown for VviGT5, which is up-regulated in a grape HY5 transient expression assay, although VviHY5 might also be involved in VviGT5 regulation. Expression of module genes was largely down-regulated during the early stages (12-24 hours post-inoculation, hpi) of leaves infected with P. viticola and E. necator, spider-mites, as well as berries infected with B. cinerea at the late stages of berry development (Fig. 4C). This strong and rapid negative response to pathogen and pest attack supports the growing evidence of the involvement of light-regulated transcriptional networks, such as the VviHY5 subnetwork, in plant-herbivore and plant-pathogen interactions.

3.6. GCC-box network highlights pathogen response and signalling in stress dataset

Although individual CRE modules may reveal important biological insights, many CRE modules often overlap as underlying genes are shared between similar as well as distinct CRE modules. In the case of similar CREs, the canonical GCC-box sequence (GGCGCC), a well-known binding site for DREB and ERF subfamilies of AP2/ERF TFs, mediates ethylene and jasmonate signaling in response to pathogens. Recent studies now show that GCC-related elements (GGCGTC and GCGGCC) are also CREs targeted by DREB and ERF TFs. In this study, GCC-box, and GCC-related CREs (GGCGTC and GCGGCC) were found to be enriched. GCNs were built for these CREs. Each of three CREs’ GCNs contained at least one module with an AP2/ERF TF: Modules I and VI for GCGGCC, Module II for GCCGTC, and Modules I, III, and IV for GCGGCC (Supplementary Table S7). Comparing those modules, the expression profiles of their genes were highly similar (Supplementary Fig. S3) and were enriched with genes pertaining to ethylene and jasmonate signaling. Furthermore, when comparing the genes involved in their networks, GCCGTC contained 82.9% unique genes, GCCGTC had 42.2% unique genes, and GCGGCC had 37.8% unique genes (Fig. 5A). Taking note of the similarity between these networks, a GCC core network (Fig. 5B) comprised of 428 genes was formed by merging the three mentioned networks. From the modules generated, two of them, Module I (Fig. 5C) and (II and Fig. 5D) contained various combinations of AP2/ERF CREs.

In Module I, five predicted grapevine AP2/ERFs (VviERF106, VIT_15s0021g01630; VviERF107, VIT_16s0013g00890; VviERF127, VIT_07s0005g03220; VviERF132, VIT_14s0081g00730; VviERF1L2, VIT_07s0005g03230) were found (Fig. 5C), all of which are homologous to group IX (B-3) AP2/ERFs in Arabidopsis. Members of group IX (B-3) AP2/ERFs are widely implicated in pathogen responses in Arabidopsis. In this module, VviERF1L2 is directly co-expressed with several genes containing GCC-related CREs including one class I chitinase (VIT_07s0005g02560) which may fulfill basic defense-related roles in grapevine, an Arabidopsis homolog of 12-oxophytodienoate reductase (VviOPR1, VIT_18s0041g02070) involved in the detoxification of harmful oxylipins accumulated during pathogen attack or wounding and potentially JA biosynthesis, and a protein kinase gene (VIT_13s0019g01080) closely related to Arabidopsis BOTRYTIS-INDUCED KINASE1 involved in the signaling cascade for defense response against pathogens. Other basic defense-related genes such as a pathogenesis-related protein (VIT_05s0077g01690) containing a GCC-box CRE is connected with VviERF132. Phenylpropanoid pathway genes encoding stilbene synthases (VIT_10s0042g00850 and VIT_10s0042g00860) and a cinnamyl-CoA reductase (VIT_14s0066g01150) involved in the biosynthesis of stilbene and lignin compounds implicated in grapevine biotic stress response, were directly connected with VviERF106 and VviERF127. One ethylene biosynthesis gene encoding 1-aminoacyclopropane-1-carboxylic acid oxidase (VIT_12s0059g01380) was directly co-expressed with VviERF107. The gene expression profile of module genes involved in hormone biosynthesis (ethylene and JA), signalling, secondary metabolism, and transcriptional regulation, were induced by B. cinerea infection in berries (24 hpi), leaves upon herbivory attack by spider mite, and leaves with powdery mildew (5 days post-infection) in the susceptible genotype Carignan (Fig. 5E). This observation supports the participation of module genes and multiple GCC-related CREs in grapevine response to pathogens (necrotrophic and biotrophic) and insect herbivory, whereby activation of inducible defense mechanism via ethylene and JA pathways might confer basal defense against these biotic stressors.
The presence of the GCC-box CRE in promoters of VviWRKY33 and direct connectivity with VviERF2 highlights a potential conserved regulation with that of Arabidopsis, where VviWRKY33 might be a target of VviERF2 in grapevines. Arabidopsis WRKY40 also contributes to broad resistance against necrotrophs, biotrophs, and hemibiotrophs, through the modulation of JA and SA crosstalk when facing different pathogen lifestyles. In addition, a NAC TF gene homologous to Arabidopsis ATAF1 (ATAF1, VIT_07s0031g02610), was directly connected with VviERF2. Arabidopsis ATAF1 is involved in both JA and ABA crosstalk and has dual roles in abiotic and biotic stress. Together, the co-expression network surrounding VviERF2, VviERF017, and VviRRTF1, may be implicated in hormonal crosstalk during plant immune response and abiotic stress through cooperative action with WRKY and NAC TFs. Both abiotic and biotic transcriptional networks are known to be mediated by AP2/ERF, WRKY, and NAC TFs regulons in plants. One pathogen-related secretory pathway gene encoding grapevine PENETRATION1 (VviPEN1, cinerea). The presence of the GCC-box CRE in promoters of VviWRKY33 and direct connectivity with VviERF2 highlights a potential conserved regulation with that of Arabidopsis, where VviWRKY33 might be a target of VviERF2 in grapevines. Arabidopsis WRKY40 also contributes to broad resistance against necrotrophs, biotrophs, and hemibiotrophs, through the modulation of JA and SA crosstalk when facing different pathogen lifestyles. In addition, a NAC TF gene homologous to Arabidopsis ATAF1 (ATAF1, VIT_07s0031g02610), was directly connected with VviERF2. Arabidopsis ATAF1 is involved in both JA and ABA crosstalk and has dual roles in abiotic and biotic stress. Together, the co-expression network surrounding VviERF2, VviERF017, and VviRRTF1, may be implicated in hormonal crosstalk during plant immune response and abiotic stress through cooperative action with WRKY and NAC TFs. Both abiotic and biotic transcriptional networks are known to be mediated by AP2/ERF, WRKY, and NAC TFs regulons in plants. One pathogen-related secretory pathway gene encoding grapevine PENETRATION1 (VviPEN1, cinerea).
VIT_08s0032g01150) was directly co-expressed with VviERF2 and VviRRTFI. The PEN1-mediated secretory pathway is an important defense strategy against powdery mildew, mediating penetration resistance. The gene expression profile of the GCC module II subnetwork is similar to module I, concordant with the role of module genes involved in immunity against necrotrophs as observed in Arabidopsis (Fig. 5E). Striking differences were observed especially in grapevine leaves infected with *P. viticola*, as many module genes were down-regulated, potentially highlighting antagonism between JA and SA signaling pathways. Down-regulation of JA signaling might alleviate repression of SA signaling, and activate early defense-JA and SA signaling pathway. Down-regulation of JA signaling were down-regulated, potentially highlighting antagonism between Arabidopsis (Fig. 5E). Striking differences were observed especially genes involved in immunity against necrotrophs as observed in Arabidopsis, both NST1 and MYB83 are master switches for secondary cell wall development, albeit NST1 function as the primary master switch while MYB83 is downstream in the hierarchy.

### Table 2. MapMan enrichment of selected MYB and GLK submodules in the development dataset

<table>
<thead>
<tr>
<th>CRE</th>
<th>Module</th>
<th>x</th>
<th>n</th>
<th>k</th>
<th>P value</th>
<th>FDR</th>
<th>Mapman term</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGATTCT</td>
<td>I</td>
<td>7</td>
<td>15</td>
<td>27</td>
<td>1.47E-13</td>
<td>9.98E-11</td>
<td>PS.lightreaction</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>7</td>
<td>17</td>
<td>148</td>
<td>1.09E-05</td>
<td>3.71E-03</td>
<td>Cell wall.degradation</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>2</td>
<td>34</td>
<td>6</td>
<td>1.94E-05</td>
<td>2.63E-02</td>
<td>Hormone metabolism.jasmonate</td>
</tr>
<tr>
<td></td>
<td>VI</td>
<td>6</td>
<td>43</td>
<td>20</td>
<td>2.58E-13</td>
<td>1.17E-10</td>
<td>Cell wall.hemicellulose synthesis</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>3</td>
<td>43</td>
<td>47</td>
<td>3.15E-07</td>
<td>1.07E-04</td>
<td>Cell wall.cellulose synthesis</td>
</tr>
<tr>
<td></td>
<td>VII</td>
<td>7</td>
<td>15</td>
<td>498</td>
<td>2.15E-09</td>
<td>1.46E-06</td>
<td>Stress.abiotic</td>
</tr>
<tr>
<td></td>
<td>VIII</td>
<td>9</td>
<td>9</td>
<td>268</td>
<td>3.77E-19</td>
<td>5.12E-16</td>
<td>Secondary metabolism.isoprenoids</td>
</tr>
<tr>
<td></td>
<td>IX</td>
<td>30</td>
<td>78</td>
<td>389</td>
<td>2.66E-36</td>
<td>9.04E-34</td>
<td>Secondary metabolism.flavonoids</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>3</td>
<td>6</td>
<td>512</td>
<td>9.94E-05</td>
<td>4.50E-02</td>
<td>DNA</td>
</tr>
<tr>
<td></td>
<td>XV</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>6.46E-07</td>
<td>8.78E-04</td>
<td>Lipid metabolism.TAG synthesis</td>
</tr>
</tbody>
</table>

x, overlap size; n, query size; and k, term size represent the contingency tables for MapMan enrichment calculations.

3.7. Different cell wall remodelling genes are enriched in subnetworks of the CCWACC development-specific network

Differences in biological function enrichments were also observed for many CRE submodules constructed from the developmental and stress-specific GCNs. For example, CCWACC commonly referred to as an AC-element, is one of the three main R2R3-MYB binding sites. It showed bona fide CRE properties (CF 1.90, Z-score 3.61), and FE 1.86 in stress-responsive, and CF 1.87, Z-score 3.10, and FE 1.87 in development-dynamic genes, and was highly enriched for biological functions in both sets of genes. However, in the developmental dataset greater enrichment for biological pathways, especially cell wall metabolism, was observed compared with stress. Therefore, focus will be given to the development-specific CCWACC network.

Two modules in the CCWACC network were enriched with distinct cell wall remodeling gene families (Table 2). The Module I (Fig. 6A) contained genes related to primary cell wall-related functions, while the Module VI (Fig. 6B) contained genes related to secondary cell wall metabolism. One MYB gene in module I (VviMYB91A, VIT_08s0007g00410) encodes an orthologue of Arabidopsis MYB91/ASYMMETRIC LEAVES 1. In Arabidopsis, MYB91/ASYMMETRIC LEAVES 1 regulates the development of several vegetative and reproductive organs including leaves and flowers. Additionally, a SQUAMOSA PROMOTER-BINDING PROTEIN 8 gene (VviSPL8, VIT_15s0021g02300), was shown to regulate the floral and seed development in Arabidopsis, was directly connected to VviMYB91A. Many cell wall remodelling genes involved in degradation and modification containing CCWACC CREs in their promoter were also present, indirectly and directly connected to VviMYB91A and VviSPL8, respectively. Examples include one gene encoding expansin B (VviEXPB2, VIT_12s0059g00190), pectin methyltransferase (VviPME3, VIT_03s0017g01950), two predicted polygalacturonases (VviPG, VIT_05s0020g0420 and VIT_15s0046g02000), and orthologues of Arabidopsis Beta-xylosidase 4 (VviXYL4, VIT_12s0035g01180 and VIT_12s0121g00230). Module genes were highly expressed during the early growth stages and decreased during development of various vegetative (e.g. rachis, tendril, leaf, stem, and seedling) and in reproductive (e.g. inflorescence, flowering, berry skin, and seed) organs (Fig. 6D). This observation reveals a strong coordinated regulation of module genes and TFs (e.g. VviMYB91A and VviSPL8) which might be critical for precise growth control during early stages of development.

Module VI contained a homolog of Arabidopsis MYB83 (VvitMYB83, VIT_06s0004g02110), which is widely implicated in the regulation of secondary wall biosynthesis components. According to our map, genes directly connected to VvitMYB83 are widely implicated in secondary cell wall biosynthesis and deposition. These include a fasciclin arabinoxylan-protein (VviFLA11, VIT_06s0004g03050), a glucuronontransferase (VviGUX2, VIT_18s0001g10580), a galacturonosytransferase (VviGUA2, VIT_04s0004g01500), a COBRA protein (VviCOBL4, VIT_14s0038g01150), one cellulose synthase (VvitCESA8, VIT_06s0003g01560), and two laccase (VvitLAC17, VIT_05s0003g01560, and VvitLAC17, VIT_15s0019g0190, VIT_13s0019g02160) genes. Interestingly, one NAC TF (VvitNST1, VIT_02s0025g02710), homolog to Arabidopsis NST1, was directly connected with VvitMYB83. In Arabidopsis, both NST1 and MYB83 are master switches for secondary cell wall development, albeit NST1 function as the primary master switch while MYB83 is downstream in the hierarchy.
3.8. AGATTCT developmental network module is specifically enriched in photosynthesis-related genes

Although CREs and CRE-driven subnetworks, such as those previously discussed, show that strong positional bias towards the TSS and FE scores provide reasonable proxies for identifying bona fide CREs and biologically relevant modules, some CREs and CRE-driven subnetworks lack these properties. For example, the 7-mer CRE AGATTCT (CF 2.69, Z-score 0.13, and FE 6.91 in the developmental dataset), whose network contains a G2-like TF, DNA-binding specificities of lesser known plant TF families such as G2-like, a subclass of GARP proteins, have been recently characterized. The G2-like TF, GLK, is implicated in photosynthesis and chloroplast development in many plants and is a key factor regulating the latitudinal gradient of ripening and various quality traits in fruits. Arabidopsis GLK1 recognizes the AGATTCT CRE sequence. The AGATTCT network in the developmental dataset contained 113 genes and was highly enriched for many photosynthesis-related genes, especially in modules I and II (Table 2) suggesting that GLK-regulated gene networks are highly conserved across plants including grapevine. Present in module I (Fig. 6C) are VviGLK1, the homolog of Arabidopsis GLK1 (VviGLK1, VIT_12s0028g03100), and its likely target genes which encode for photosystem I polypeptide subunits (VviPSAN, VIT_04s0044g01410; VviPSAF, VIT_00s0125g00280; VviPSAD-2, VIT_05s0020g03180) and light harvesting complexes (VviCBAB1-1, VIT_10s0030g02890; VviCBAB1-2, VIT_10s0030g02900; VviLBh6, VIT_12s0055g01110). One gene involved in cellular redox homeostasis (VviCDSP32, VIT_18s0001g00820) was also implicated in this module. The expression survey across the developmental dataset revealed that module I genes are highly expressed in immature organs of the berry pericarp, flesh, skin, and seed, as well as in rachis, buds, and leaves, and gradually decrease their expression as development of each organ progresses (Fig. 6D). Therefore, the consistent down-regulation of module genes may reflect an active role of photosynthetic genes and VviGLK1 in the immature-to-mature phase transition of grapevine organs, similar to the observations in other plants. Additionally in grape berries, these module genes may be critical for the establishment of differing ripening gradients during the onset of ripening, as seen with the role of GLK1 in tomatoes.

3.9. Comparison between grapevine and Arabidopsis motif-driven GCNs

Compared with other plants, CRE-driven GCNs have only been elucidated in Arabidopsis; several of which include the G-box, MYB (CCWACC), WRKY (KTTGACY), and the site II element (TGGGCG) CRE modules. Based on the conserved G-box and CCWACC CRE, which are also evaluated in this study, many enriched (GO enrichment) biological pathways in Arabidopsis are shared with those identified in this study. For example, highly enriched terms related to photosynthesis, hormone (ethylene and ABA) metabolism, and stress responses were common in G-box CRE modules between grapevine (Supplementary Table S3) and Arabidopsis. Similarly, secondary metabolic pathways, especially related to flavonoid and phenylpropanoid metabolism, cell wall, and stress responses, were common for the CCWACC (MYB) CRE module. To determine if other grapevine CRE modules harbor unique and/or common features at a larger scale, we expanded the search in Arabidopsis by constructing CRE modules (P < 0.01) using the Arabidopsis graphical Gaussian model gene network with a partial correlation threshold > 0.05 (Ma et al. (Supplementary Table S9). We then determined their biological pathway (Mapman) enrichment profiles (Supplementary Table S10) and performed the comparison with CRE modules from grapevine. In addition to the G-box and CCWACC CRE, some grapevine CRE modules share a high degree of similar gene enrichments with Arabidopsis CRE modules which includes AGCGGC (ABRE), TGGCCG, TTTCCTT, and YAACKG (MYB) (Supplementary Table S10), suggesting similar properties of gene expression and gene networks between grapevine and Arabidopsis. However, others share a small proportion of enriched pathways and/or have entirely unique enriched terms such as GCCGCC, RCCGAC, TTGCGT (bZIP), TTGACC (W-box), and AP2/ERF (ACCGACA).
especially cold, osmotic stress, and salt, but also to UV-B, drought, and heat stress albeit at a lesser extent (Supplementary Table S11A). Biotic stress pathways are also highly modulated in the G-box module I of Arabidopsis and a greater proportion of orthologues are down-regulated against different biotic stress including B. cinerea and powdery mildew, as seen in grapevine (Supplementary Table S11B). In addition, in module I (Supplementary Table S11C) and II (Supplementary Table S11D) of the stress-specific GCC-core module, there are 47 (of 102) and 66 (of 94) grapevine genes which have Arabidopsis orthologues, respectively. Like grapevine, Arabidopsis orthologues for these genes were highly induced upon a wide variety elicitors and plant–pathogen treatments. Although similar gene expression responses were observed for B. cinerea and powdery mildew infection for grapevine and Arabidopsis, an opposite response (induction) was observed for Arabidopsis infection with H. arabidopsidis, the causal agent of downy mildew (Supplementary Table S11C and D), compared with grapevine (Fig. 5E). This highlights potential differences in the transcriptional networks of plant immunity employed by the two species that may involve differing CRE regulation. Nonetheless, the strong and consistent modulation of gene orthologues of CRE modules between grapevine and Arabidopsis suggest a general conservation in promoter architecture, gene expression dynamics, and gene regulatory networks.

4. Conclusion

In this study, we have elucidated the properties of a large selection of known plant CREs in grapevine promoters at the genome-scale. We observed that many CREs fulfill the widely-adopted criteria defining authentic plant CREs. We corroborated these findings through extensive biological pathway enrichment analysis suggesting the characterized CRE structure has strong biological relevance. Together, these assessments have elucidated many potential targets of a large repertoire of grapevine TFs for the first time. Large-scale transcriptome analysis using the developmental atlas dataset and a selection of stress-related RNA-seq datasets demonstrated that the distribution of CREs in gene promoters explains close to half of the DE across the various abiotic and biotic stresses, and developmental comparisons evaluated. However, several exceptions exist where well-characterized plant CREs do not possess authentic CRE properties and/or strong enrichment, highlighting the potential diversification of CREs in grapevine promoters and divergence of transcriptional regulatory pathways. Therefore, additional scanning approaches or other statistical models, such as scans using position weight matrices or all possible k-mer nucleotide combinations, might be necessary to capture the full spectrum of grapevine CREs and their biological functions. Nonetheless, by integrating condition-specific GCC, CRE enrichment, and gene expression profiles, numerous condition-specific CRE-driven modules (and submodules) were discovered, providing additional information on the coordinated regulation of submodule genes. Although some of these genes may not contain the given CRE, they are implicated in similar functions, specificity of stress responses, and/or gene expression pattern in tissues/organisms within and between CRE-driven submodules. This integrated approach uncovered biologically meaningful submodules in a more robust fashion than just utilizing the distribution of CREs alone. This is exemplified by the modules that corroborated known roles in drought response, pathogen defense, cell wall metabolism, and fruit ripening in grapevine and in other plant species. Some modules reveal novel regulatory functions, although further functional validation will be necessary to confirm the transcriptional control exerted by the large TF repertoire in grapevines. Application of these approaches is not just limited to grapevines but to other sequenced plant genomes, especially crop species, permitting a greater understanding of the evolution and conservation of plant CREs and transcriptional regulatory pathways.

Acknowledgements

We thank the grapevine research community for making various RNA-seq and microarray datasets publicly available.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study was funded by the University of British Columbia (10R18459), Genome British Columbia (10R21188), and the Natural Sciences and Engineering Research Council of Canada (10R23082).

References


