TITLE: TEA-GCN: Constructing Accurate, Species-comparable, and Explainable Gene Regulatory and Functional Networks via Condition-Specific Ensemble Co-expression

AUTHORS: Peng Ken Lim[1*], Ruoxi Wang[1], Shan Chun Lim[1], Jenet Princy Antony Velankanni[1], Marek Mutwil[1,2*]

[1]*School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore*

[2]*Department of Plant & Environmental Sciences, University of Copenhagen, Frederiksberg C 1871, Denmark*

*Corresponding authors:

Marek Mutwil

Department of Plant & Environmental Sciences,

University of Copenhagen,

Frederiksberg C 1871,

Denmark

Email: mutwil@plen.ku.dk

Peng Ken Lim

School of Biological Sciences,

Nanyang Technological University,60 Nanyang Drive,

637551, Singapore,

Singapore

Email: pengkenlim.sbs@gmail.com , pengken001@e.ntu.edu.sg (Institutional)

Abstract

Gene co-expression networks (GCNs) are widely used for gene function prediction, yet their performance often suffers from low dataset quality. To address this problem, we have developed TEA-GCN (Two-Tier Ensemble Aggregation-GCN), a novel method that constructs highly-performing ensemble GCNs by leveraging unsupervised transcriptomic data partitioning and multi-metric co-expression scoring. Benchmarking over 450,000 public RNA-seq samples across *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Homo sapiens*, and nine additional angiosperms, we show that TEA-GCN consistently outperforms the state-of-the-art in predicting gene functions, recovering transcription factor-target relationships, and inferring gene regulatory networks. TEA-GCN uses natural language processing to provide an unprecedented level of explainability by identifying co-expression relationships unique to specific tissues or conditions. Additionally, TEA-GCNs exhibit enhanced cross-species conservation, improving comparative analyses across diverse plants. We provide both an open-source pipeline (https://github.com/pengkenlim/TEA-GCN) and a user-friendly Plant-GCN web resource for ten Angiosperm species (https://plantgcn.connectome.tools/), supporting widespread application of this framework in gene function prediction and comparative transcriptomics.

Introduction

Gene co-expression networks (GCNs) play a pivotal role in revealing co-regulatory and co-functional relationships between genes[1–3]. This concept, rooted in the observation that functionally related genes often exhibit coordinated expression patterns, has proven invaluable in predicting gene function and uncovering modules of co-regulated genes[2–5]. Using GCNs to uncover biological insights becomes especially important in non-model species, where experimental validation of gene function is scarce[6,7]. The rapidly declining costs of RNA-seq, coupled with the fast rate at which new genomes are published, has led to an explosion of publicly available transcriptomic data, particularly in non-model species[6–8]. These recent developments offer an unprecedented opportunity to unravel the evolutionary dynamics of gene function across diverse organisms via comparative-GCN analysis[5,6,9,10].

However, the investigative potential of GCNs is hindered by several challenges when working with public transcriptome data. The types and balance of samples in a transcriptomic dataset strongly influence the accuracy of GCNs and the biological insights they reveal [1,11,12]. For example, the overrepresentation of certain sample types (e.g., tissues, conditions) within the dataset can lead to overrepresentation of certain regulatory information [13]. Additionally, the size and heterogeneity of the transcriptome data can also have a huge impact on the resultant GCN. For example, dataset size (i.e., the number of samples) can directly affect the statistical power and, therefore, the resolving power of GCNs to differentiate authentic transcriptional co-regulation of genes from spurious correlation of expression patterns[12]. On the other hand, GCNs built from a highly heterogeneous dataset consisting of diverse sample types can excel in capturing ubiquitous gene modules but can fail to capture condition-specific regulatory relationships due to insufficient sample-type replication[13]. Thus, the variability of publicly available RNA-seq data of different species, in terms of sample quality, heterogeneity, and size, poses a substantial hurdle for constructing performant GCNs. To address these challenges, many studies have utilized highly curated public transcriptomic datasets that are sample-balanced and/or context-specific to construct well-performing GCNs[14–18]. However, these methods rely on labor-intensive sample annotation that might not be available for some species.

Recent advancements in obtaining performant GCNs involve ensemble approaches that combine multiple GCNs to enrich biological information[19–21]. In one study, gene expression omnibus (GEO) datasets of *Vitis vinifera* (Grape) were used to generate unweighted GCNs and aggregate them into an ensemble GCN with dramatically improved performance[20]. However, this

approach might not work for cases where the number of GEO datasets is limited, even for species that possess numerous public RNA-seq accessions. Another approach (henceforth referred to as the “Subbaging” method) combined weighted GCNs calculated from random subsets of principal components of the gene expression matrix to yield performant ensemble GCNs for Human, Fission Yeast, Budding Yeast, 9 other Metazoan and 11 Plant species[15,18]. According to the authors, combining GCNs generated from random subsets of the whole data achieves a sample-balancing effect that increases GCN performance[15,18]. However, the stochastic nature of random subsampling, as well as the ambiguous meaning of principal components, compromises the interpretability of the resultant ensemble GCN. In addition, both approaches only use Pearson’s correlation coefficient (PCC), which can fail to consider nonlinear co-expression relationships between genes[1,22].

In response to these challenges we have developed a novel method for GCN generation, called the Two-Tier Ensemble Aggregation GCN (TEA-GCN) that does not require annotated samples for batch correction and/or sample balancing, making it readily applicable to large public transcriptomic datasets with limited preprocessing and suitable for species with scarce sample annotation.

Results

Overview of Two-Tier Ensemble Aggregation Gene Co-expression Network (TEA-GCN) Framework

In order to create a more robust measure of co-expression, TEA-GCN combines the Coefficient and Partition aggregation in two steps. In the first step, Coefficient Aggregation combines the three coefficients in every partition, and in the second step, Partition Aggregation combines the coefficients from the different partitions (Figure 1A).

Coefficient Aggregation involves the calculation of different co-expression metrics (coefficients) between a given pair of genes and then unifying (aggregating) said metrics into one measure of co-expression strength (Figure 1A). We used the three most popular metrics: Pearson's correlation coefficient (PCC, capturing linear relationships), Spearman's rank correlation coefficient (SCC, capturing monotonic relationships)[1,23,24] and biweight midcorrelation[25] (Bicor), a coefficient employed in the widely used WGCNA R package[26] that is more resistant to outliers and noise than PCC[25] . Although several studies compare the three coefficients, there is no current consensus as to which is the most ideal for generating robust GCNs, as different coefficients excel in detecting different biological relationships[24]. Therefore, to detect more types of complex yet biologically significant co-expression relationships between genes, the TEA process uses Coefficient Aggregation to derive co-expression based on all three coefficients using the Maximum (Max) function (see Supplemental methods).

Partition Aggregation aims to capture tissue/condition-specific GCNs that might be missed if all samples are used for the network construction. The TEA-GCN framework uses K-means clustering[32], an unsupervised clustering algorithm, to partition the transcriptomic dataset into samples with similar transcriptome profiles[33] (Figure 1A). Through Partition Aggregation, the co-expression strengths between all gene pairs are aggregated across all partitions to detect tissue/condition-specific co-expression using the Rectified Average (RAvg) function, where the negative co-expression strengths from every dataset partition are converted to zero before averaging, resulting in more accurate co-expression determination (supplemental results, Figures S1 and S2). The resultant ensemble co-expression from two rounds of aggregation was then subjected to Mutual Rank Transformation (supplementary methods) to yield the final TEA-GCN.

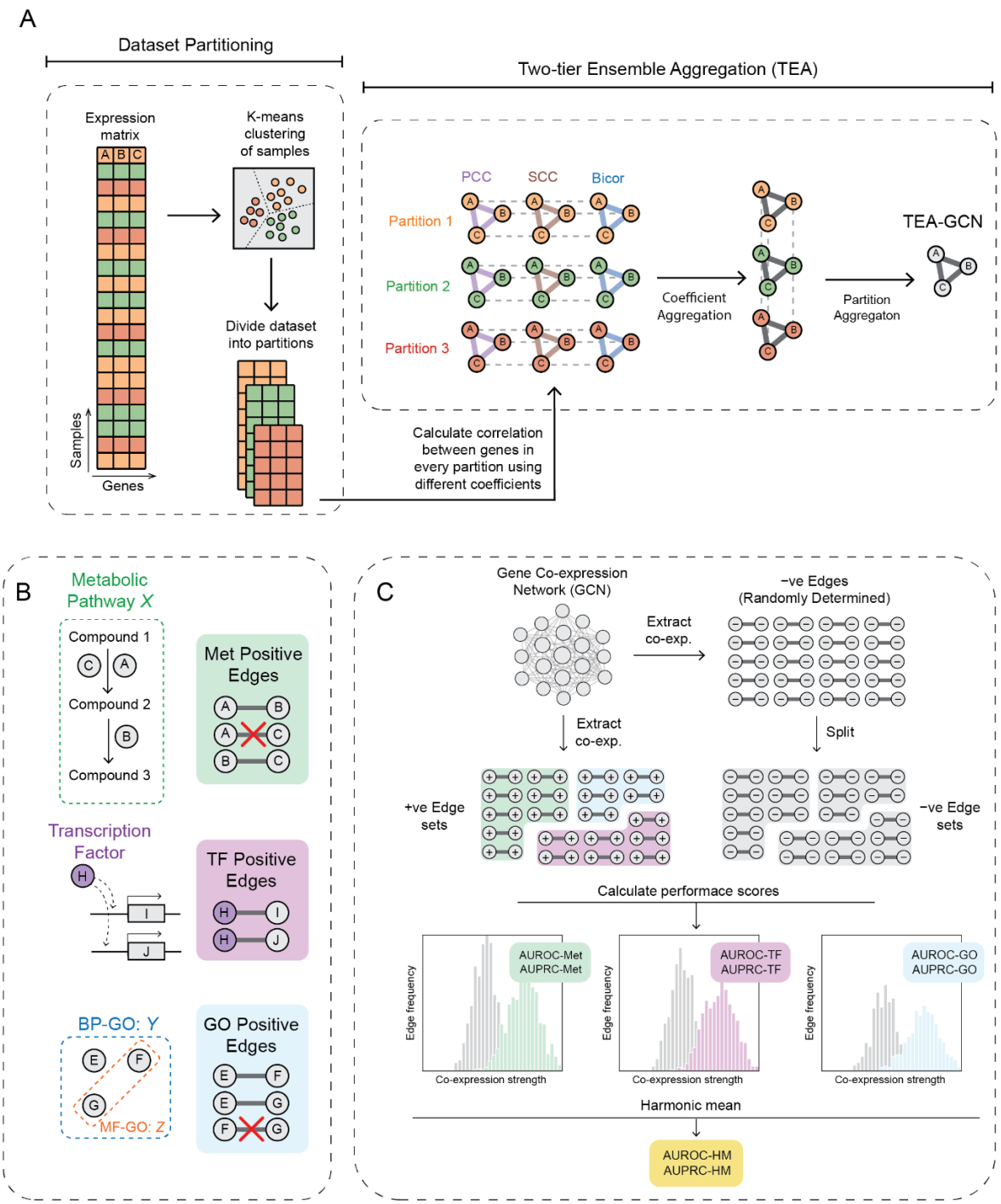


Figure 1. Overview of the Two-Tier Ensemble Aggregation Gene Co-expression Network (TEA-GCN) framework and GCN evaluation methodology. A) The processing of data from an initial gene expression matrix to the final ensemble GCN is visualized using a toy model in which there are only three genes (A, B, C) in the genome. The gene expression matrix is first partitioned via k-means clustering at a clustering granularity (k=3) that has been determined using a combination of semi-supervision and/or internal clustering evaluation metrics. For each dataset partition, GCNs are generated based on three different correlation coefficients and aggregated into one ensemble GCN. Resultant GCNs from each partition are again aggregated to yield the final ensemble TEA-GCN. GCNs with coexpression strength represented by Pearson's correlation coefficient (PCC), Spearman's rank correlation coefficient (SCC), and Biweight midcorrelation (Bicor) are denoted by purple, brown, and blue edges, respectively. The node color of GCNs

corresponds to the particular dataset partition they represent. **B)** We define three sets of positive edges connecting genes sharing different biological relationships. Positive Metabolic (Met) edges are made by pairing experimentally validated enzymes from the same pathway. Edges connecting enzymes that catalyze the same reaction step are removed from positive Met edges. Positive Transcription factor (TF) edges are generated between genes and their corresponding activator TFs. Positive Gene Ontology (GO) edges are made from pairing genes that were annotated for the same biological process (BP) GO term at the IDA (inferred from direct assay) evidence level. Edges connecting genes annotated with the same molecular function (MF)-GO term are removed from positive GO edges. **C)** To evaluate a given GCN, co-expression strengths of positive edge sets for each biological aspect are compared to co-expression strengths of randomly determined negative edges of the same size using the area-under-curve (AUC) of the receiver operator characteristics curve (ROC) and precision-recall curve (PRC). Scores calculated for each positive-negative edge set pair represent a GCN's performance to identify different biologically significant relationships between genes and can be aggregated using the harmonic mean to infer a GCN's overall performance.

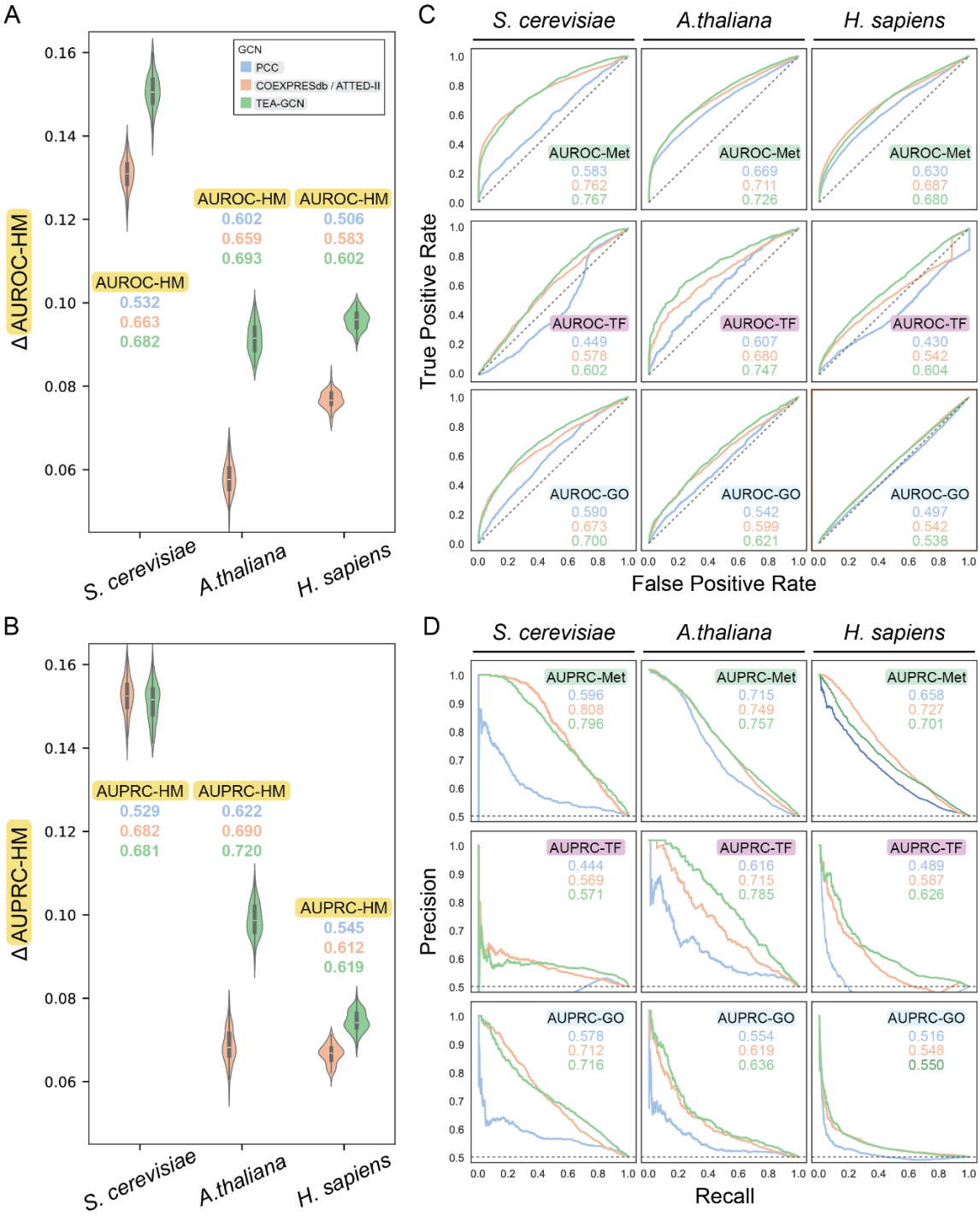
TEA-GCN achieves better overall performance than the state-of-the-art

We benchmarked TEA-GCN using 150,096, 57,759, and 95,553 public RNA-seq samples for *S. cerevisiae*, *A. thaliana*, and *H. sapiens,* respectively (see Methods)., we evaluated the ability of networks to recover experimentally-verified metabolic networks (Met) (Figure 1B; green), transcription factor-target relationships (TF) (Figure 1B; purple), and genes involved in the same biological process (GO) (Figure 1B; blue).

We found that the combination of Coefficient Aggregation and Partition Aggregation in TEA-GCN results in a dramatic increase in GCN performance for *S. cerevisiae*, *A. thaliana*, and *H. sapiens* (Supplemental Results, Figures S1 and S2). To compare its performance to the state-of-the-art approaches, we compare the performances of *S. cerevisiae*, *A. thaliana*, and *H. sapiens* GCNs from two industry-leading databases that use the 'subbagging' approach, COXPRESdb[15] (https://coxpresdb.jp/download/) and ATTED-II[18] (https://atted.jp/download/). As expected, these "Subbaging" ensemble GCNs have a better overall performance than non-ensemble PCC GCNs across all three species (ΔAUROC-HM and ΔAUPRC-HM > 0.06-0.16) (Figure 2A-B; Table S1 to 3). However, TEA-GCNs showed substantially higher performance for nearly all metrics (Figures 2A, 2B, and S3; Table S1 to 3).

To provide a more fine-grained comparison, we measured the performance differences in predicting gene regulatory relationships (TF), co-involvement in metabolic pathways (Met), and biological processes (GO-BP). TEA-GCNs were consistently more performant in retrieving the experimentally verified gene regulatory relationships (TF) across the three species (Figure 2C and 2D; Table S1 to 3). For example, *A. thaliana* TEA-GCN has a TF performance (AUROC-TF, AUPRC-TF) of (0.747, 0.785) as compared to (0.680, 0.715) for *A. thaliana* ATTED-II GCN; *H. sapiens* TEA-GCN has a TF performance of (0.604, 0.626) as compared to (0.542, 0.587) for *H. sapiens* COXPRESdb GCN (Figure 2C and 2D; Table S1 to 3). Conversely, the ATTED-II/COXPRESdb GCNs are better at identifying pathway co-involvement of enzymes for *S.*

165 *cerevisiae* and *H. sapiens*. *A. thaliana* TEA-GCN (AUROC-Met, AUPRC-Met of 0.726, 0.757) has
166 a slightly higher Met performance than *A. thaliana* ATTED-II (0.711,0.749), and better
167 performance in identifying genes sharing the same GO-BP term in *S. cerevisiae* (0.700, 0.716 vs
168 0.673, 0.712) and *A.thaliana* (0.621,0.636 vs 0.599, 0.719) (Figure 2C and 2D; Table S1 to 3).
169 We speculate that the superior performance of TEA-GCN lies in its ability to detect
170 condition/tissue-specific co-expression from individual dataset partitions. The observed TF
171 performance disparity between TEA-GCN and the "Subbaging" method is more pronounced in
172 multicellular *A. thaliana* and *H. sapiens* than in unicellular *S. cerevisiae* (Figure 2C and 2D; Table
173 S1-3), which could be attributed to the presence of more specialized transcriptional programs in
174 multicellular organisms than unicellular organisms[27,28].


A
GCN
PCC
COEXPRESdb / ATTED-II
TEA-GCN
Δ AUROC-HM
AUROC-HM
0.532
0.663
0.682
AUROC-HM
0.602
0.659
0.693
AUROC-HM
0.506
0.583
0.602
S. cerevisiae
A.thaliana
H. sapiens
B
Δ AUPRC-HM
AUPRC-HM
0.529
0.682
0.681
AUPRC-HM
0.622
0.690
0.720
AUPRC-HM
0.545
0.612
0.619
S. cerevisiae
A.thaliana
H. sapiens
C
S. cerevisiae
A.thaliana
H. sapiens
True Positive Rate
False Positive Rate
AUROC-Met
0.583
0.762
0.767
AUROC-Met
0.669
0.711
0.726
AUROC-Met
0.630
0.687
0.680
AUROC-TF
0.449
0.578
0.602
AUROC-TF
0.607
0.680
0.747
AUROC-TF
0.430
0.542
0.604
AUROC-GO
0.590
0.673
0.700
AUROC-GO
0.542
0.599
0.621
AUROC-GO
0.497
0.542
0.538
D
S. cerevisiae
A.thaliana
H. sapiens
Precision
Recall
AUPRC-Met
0.596
0.808
0.796
AUPRC-Met
0.715
0.749
0.757
AUPRC-Met
0.658
0.727
0.701
AUPRC-TF
0.444
0.569
0.571
AUPRC-TF
0.616
0.715
0.785
AUPRC-TF
0.489
0.587
0.626
AUPRC-GO
0.578
0.712
0.716
AUPRC-GO
0.554
0.619
0.636
AUPRC-GO
0.516
0.548
0.550

175

Figure 2. Benchmarking TEA-GCN across three different model organisms against "Subbaging" ensemble GCNs. Distribution of **A)** AUROC-HM performance scores and **B)** AUPRC-HM performance scores between Subbaging CGNs and TEA-GCNs across *S. cerevisiae*, *A. thaliana* and *H. sapiens.* We report the delta values of performance scores (i.e., Δ AUROC-HM and Δ AUPRC-HM) used to compare GCNs on a unified axis. The delta values were calculated by subtracting performance scores from the median performance score of the baseline non-ensemble PCC network of respective organisms. Median performance scores of PCC (blue) networks, Subbaging CGNs (orange), and TEA-GCNs (green) are indicated. **C)** Receiver operating characteristic curves (ROCs) and **D)** Precision-recall curves (PRCs) correspond to median performance scores for each biological aspect and organism. Across all panels, performance information relating to the PCC network, Subbaging CGN, and TEA-GCN is represented in blue, orange, and green, respectively.

TEA-GCN captures co-expression between experimentally validated enzymes in *Arabidopsis thaliana* metabolic pathways

To characterize how well the TEA-GCN framework captures co-expression between enzymes of the same pathway, we compared the co-expression strengths of Met positive edges in *A. thaliana* TEA-GCN and ATTED-II GCN. We define Met positive edges as edges that connect experimentally validated enzymes within individual PlantCyc pathways[29]. For a fair comparison of co-expression strengths across the two CGNs, we used standardized co-expression strengths (***Z***(Co-exp.)), which measure the co-expression strengths of each edge relative to the co-expression strengths of all edges in the GCN (see Methods).

A. thaliana TEA-GCN performed better in identifying Met positive edges than the ATTED-II GCN across all major pathway classes(Figure 3A; Grey section). The only exception was the "Generation of Precursor Metabolites & Energy" class, where performance between the two GCNs is nearly identical (Figure 3A; Table S4), irrespective of co-expression thresholds (i.e., ***Z***(Co-exp.) ≥ 0.5 and ***Z***(Co-exp.) ≥ 1). TEA-GCN excelled in the "Metabolic Clusters" and "Biosynthesis" classes, where the percentage of positive edges at both thresholds is more than 10% higher than in ATTED-II GCN (Figure 3A; Table S4). The "Generation of Precursor Metabolites & Energy" class consists of conserved pathways such as the TCA cycle, glycolysis, respiration, and photosynthesis that are ubiquitously expressed and not likely to be condition/tissue-specific, which could explain the similarity of performances in both GCNs. Conversely, "Biosynthesis" and "Detoxification" classes, which tend to comprise specialized pathways that are lineage-specific and/or condition/tissue-specific[30–35], have more co-expressed edges in TEA-GCN than ATTED-II GCN at both thresholds. These observations are compatible with our hypothesis that the good performance of TEA-GCN is due to its ability to capture condition/tissue-specific co-expression.

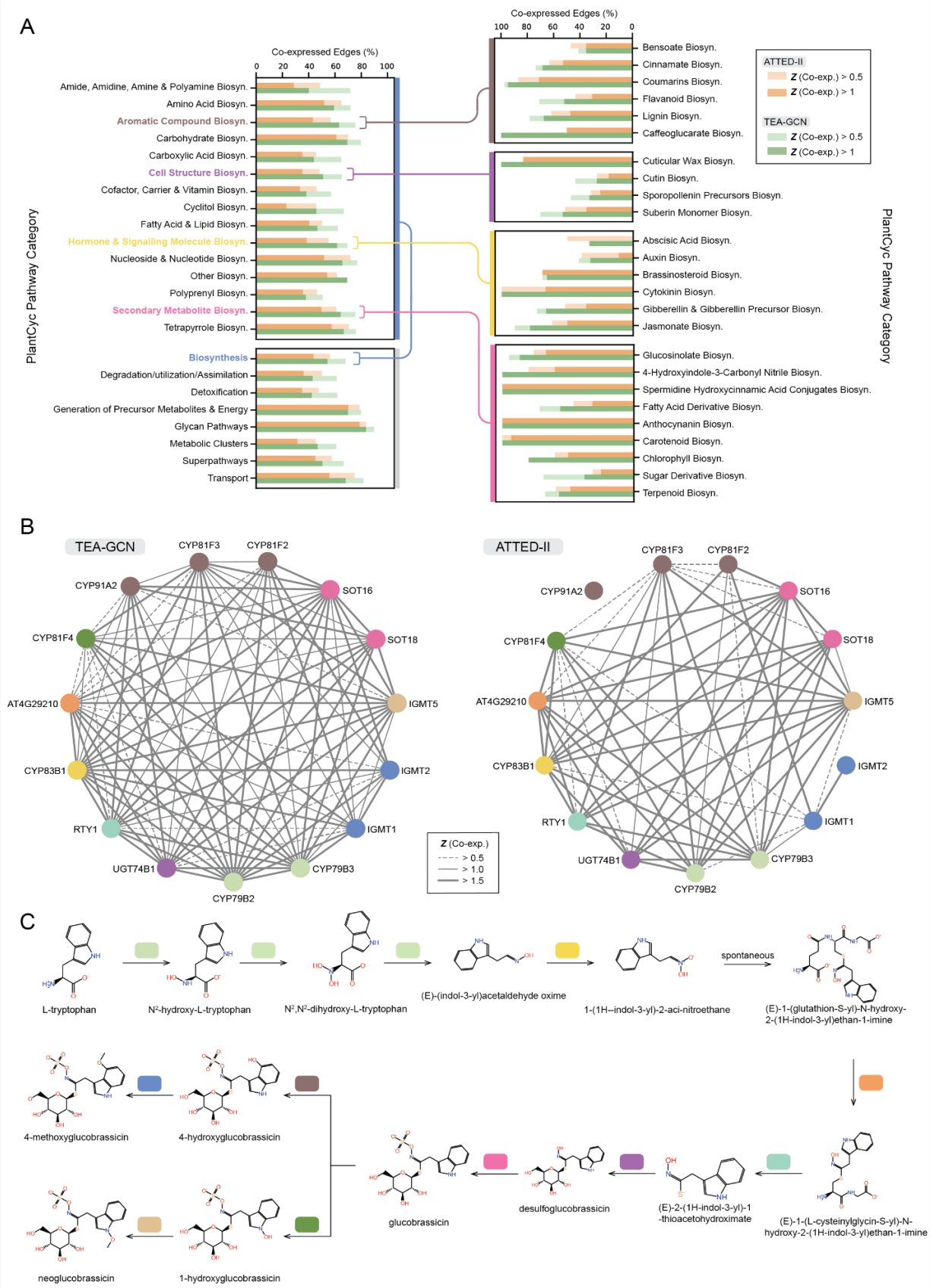


Figure 3. Characterizing TEA-GCN's ability to capture metabolic pathways in *A. thaliana*. A) Proportion of Met positive edges

of different PlantCyc pathways that are co-expressed in *A. thaliana* TEA-GCN and ATTED-II GCN at two (>0.5 and >=1) co-expression thresholds. Met positive edges connect experimentally validated enzymes of different reaction steps. Grey-labeled plot corresponds to first-level ontological classes and compares major pathway categories. The blue-labeled plot breaks down second-level ontological classes within "Biosynthesis", of which third-level ontological classes consisting of "Aromatic Compound Biosynthesis", "Cell Structure Biosynthesis", "Hormone and Signalling Molecule Biosynthesis", and "Secondary Metabolite Biosynthesis" were further broken down in the brown, purple, yellow and pink plots, respectively. **B)** Co-expression between enzymes of "Glucosinolate Biosynthesis from Tryptophan" PlantCyc pathway (PWY-601) in *A. thaliana* TEA-GCN and ATTED-II GCN. Nodes represent experimentally validated enzymes according to PlantCyc and are labeled with gene symbols (if available) or gene identifiers. The thickness of grey edges connecting nodes corresponds to standardized co-expression strengths between enzymes in the respective GCNs. **C)** Reaction steps in Glucosinolate Biosynthesis from Tryptophan PWY-601. Colored blocks in the middle of reaction steps annotate enzyme nodes in panel B based on the reaction step in which they participate.

To better compare the *A. thaliana* GCNs of TEA-GCN and ATTED-II, we investigated the different sub-classes within the "Biosynthesis" class. TEA-GCN identified at least 10% more co-expressed edges than ATTED-II for "Hormone and Signaling Molecule Biosynthesis", "Cyclitol Biosynthesis", "Aromatic Compound Biosynthesis", "Cell Structure Biosynthesis", "Secondary Metabolite Biosynthesis", "Amide, Amidine, Amine, and Polyamine Biosynthesis", at both co-expression thresholds (Figure 3A, Blue). We observed an extra 23% of positive met edges belonging to "Hormone and Signaling Molecule Biosynthesis" for TEA-GCN (61.5%; 123 out of 200) compared to ATTED-II GCN (38.5%; 77 out of 200) (Figure 3A, blue; Table S4) at the stricter co-expression threshold of ***Z***(Co-exp.) ≥ 1. This observation is largely attributed to the "Gibberellin and Gibberellin Precursor Biosynthesis" and "Jasmonate Biosynthesis" (Figure 3A, yellow) that collectively account for more than half of the edges belonging to "Hormone and Signaling Molecule Biosynthesis" (Table S4). Given that the biosynthesis of hormones is known to be highly localized to certain tissues at specific stages of development or environmental conditions[36–38], we attribute TEA-GCN's higher performance to an ability to capture condition-specific co-expression. Other notable classes where TEA-GCN excels include "Fatty Acid Derivative Biosynthesis" (TEA-GCN: 55.97% and ATTED-II: 31.45%, strict threshold; 159 total edges), "Flavonoid Biosynthesis" (TEA-GCN: 51.97% and ATTED-II: 30.71%, strict threshold; 127 total edges), "Lignin Biosynthesis" (TEA-GCN: 67.48% and ATTED-II: 47.15%, strict threshold; 123 total edges) and "Glucosinolate Biosynthesis" (TEA-GCN: 86.74% and ATTED-II: 66.85%, strict threshold; 181 total edges) (Figure 3A; Table S4).

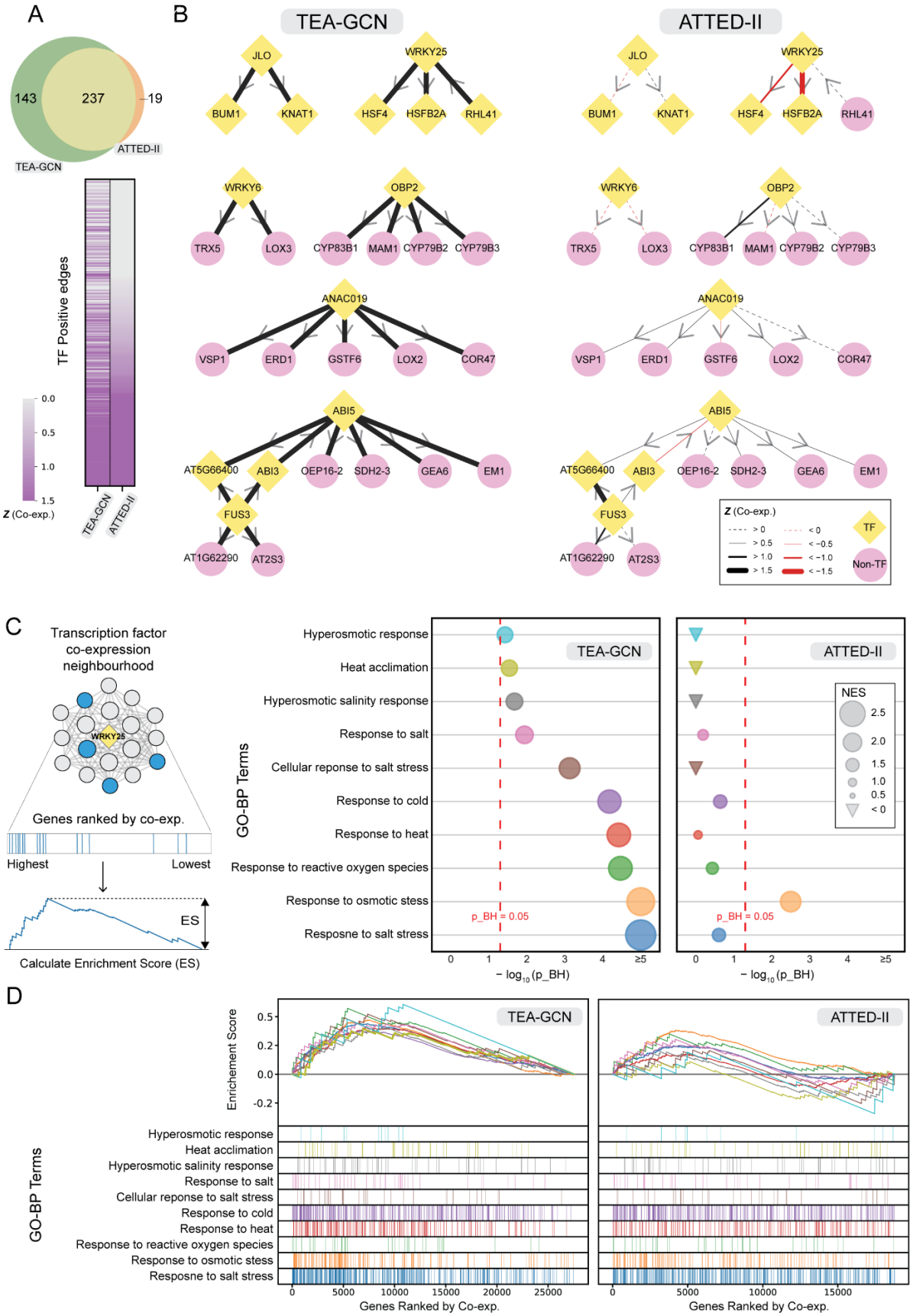
To exemplify the utility of TEA-GCN in predicting pathway memberships of enzymes, we compared the co-expression strengths between genes that code for experimentally validated enzymes participating in the "Glucosinolate Biosynthesis from Tryptophan" PlantCyc pathway (henceforth referred to by its corresponding PlantCyc pathway accession, PWY-601; Figure 3B and C). There is significant research interest in transgenically modifying the level of specific glucosinolates, which can be beneficial or toxic when consumed[39–41], in Brassicaceae food crops[42], which relies on elucidating their biosynthetic enzymes.

251 Experimentally validated enzymes of PWY-601 tended to be co-expressed with more
252 experimentally validated enzymes from the same pathway in *A. thaliana* TEA-GCN than in
253 ATTED-II GCN (Figure 3B). Furthermore, genes coding for enzymes involved in the derivatization
254 of 3-IMG to 4-MoIMG (i.e., *IGMT1*, *IGMT2*, *CYP91A2*, *CYP81F3*, and *CYP81F2*) have notable
255 lower co-expression strengths with other PWY-601 enzymes in ATTED-II GCN, as opposed to
256 the more connected PWY-601 network of TEA-GCN (Figure 3B and C). This lack of co-expression
257 with other enzymes is especially prominent for *CYP91A2* and *IGMT2*, which codes for the
258 monooxygenase and methyltransferase that catalyze the hydroxylation of 3-IMG to 4-
259 hydroxylglucobrassicin (4-OhIMG) and the methylation of 4-OhIMG to 4-MoIMG, respectively[43];
260 *CYP91A2* is not co-expressed with any genes in PWY-601, while *IGMT2* is only connected to
261 *IGMT1* which codes for a methyltransferase catalyzing the same reaction step (Figure 3B). All in
262 all, we have shown that TEA-GCN can better capture co-expression between genes coding for
263 enzymes within specialized metabolic pathways than the state-of-the-art "Subbagging" method,
264 possibly due to TEA-GCN's ability to detect tissue/conditional-specific gene co-expression.
265

266 **TEA-GCN can accurately reveal condition/tissue-specific gene regulatory networks**

267 Our analysis revealed that TEA-GCN is particularly performant in revealing transcription factors
268 and their gene targets (Figures 2C and D; TF aspect). While 35.2% of TF positive edges were
269 co-expressed in both GCNs (237 out of 673, Figure 4A), 21.2% (143 out of 673, Figure 4A) were
270 found only in TEA-GCN. Conversely, only 19 TF positive edges were uniquely co-expressed in
271 ATTED-II GCN (2.8%; Figure 4A). Thus, TEA-GCN contains nearly all edges found in ATTED-II
272 and can identify additional positive edges.
273 To further characterize the difference between the gene regulatory networks identified by
274 TEA-GCN and ATTED-II, we analyzed transcription factors where TEA-GCN, but not ATTED-II,
275 showed high co-expression (***Z***(Co-exp.) >1.5) with their transactivation gene targets (Figure 4B).
276 The transcription factor genes identified in this manner are known to be highly tissue/condition-
277 specific in their expression, which further supports our hypothesis that TEA-GCN excels in
278 detecting tissue/condition-specific co-expression. To exemplify, *JAGGED LATERAL ORGANS*'s
279 (*JLO*) expression is specific to embryos[44], while *ABA INSENSITIVE 5*'s (*ABI5*) and *FUSCA3*'s
280 (*FUS3*) expression are seed-specific[45,46]. *OBF-BINDING PROTEIN 2*'s (*OBP2*) expression is
281 induced by wounding and phytohormone Methyl jasmonate (MeJA)[47] , *NAC DOMAIN*
282 *CONTAINING PROTEIN 19*'s (*ANAC019*) expression is induced by drought, high-salt stress and
283 phytohormone abscisic acid[48], and *WRKY6*'s expression is induced by phosphate-starvation[49].

284 Curiously, *WRKY25* is found by ATTED-II GCN to be negatively co-expressed with its
285 transactivation targets of *HEAT SHOCK FACTOR 4* (*HSF4*) and *HEAT SHOCK*
286 *TRANSCRIPTION FACTOR B2A* (*HSFB2A*)[50] (Figure 4B). *WRKY25* is involved in responses
287 against a multitude of abiotic stresses such as heat, cold, osmotic, and high-salt conditions[50–53].
288 Given the multimodal functions of *WRKY25* and its participation in various abiotic responses, we
289 speculate that its mode of action might be condition-specific, which might be the reason why its
290 co-expression with *HSF4* and *HSFB2A* were not detected in ATTED-II GCN.
291 Next, to investigate whether the identified transcription factors (Figure 4B) are co-
292 expressed with biologically relevant genes, we compared the functional enrichment of genes co-
293 expressed with the selected transcription factors for *A. thaliana* TEA-GCN and ATTED-II GCN.
294 To this end, we encapsulated the co-expression neighborhoods of transcription factors as lists of
295 genes ranked by co-expression and used them as input for Gene Set Enrichment Analysis
296 (GSEA) using GO-BP gene sets (Figure 4C). We found that the functional enrichment profiles of
297 co-expression neighborhoods of selected transcription factors are significant for many of their
298 known biological functions in TEA-GCN, but not for ATTED-II (Figure 4C and D, Figure S4). For
299 example, genes involved in “Hyperosmotic response” (GO:0006972), “Hyperosmotic salinity
300 response” (GO:0042538), “Response to salt” (GO:1902074), “Cellular response to salt stress”
301 (GO:0071472), “Response to salt stress” (GO:0009651), “Heat acclimation” (GO:0010286),
302 “Response to heat” (GO:0009408), “Response to cold” (GO:0009409) and “Response to reactive
303 oxygen species” (GO:0000302) are enriched ($p_BH \leq 0.05$) in the *WRKY25* co-expression
304 neighborhood of TEA-GCN but not that of ATTED-II GCN (Figure 4D and E). Although both
305 *WRKY25* co-expression neighborhoods from *A. thaliana* TEA-GCN and ATTED-II GCN are
306 enriched for genes involved in “Response to osmotic stress” (GO:0006970), the former has a
307 higher magnitude of enrichment according to NES scores (Figure 4C and D). Thus, in addition to
308 showcasing the robust gene function prediction abilities of the TEA-GCN, we have exemplified its
309 potential to predict functions of genes involved in multiple biological processes that may be elusive
310 to traditional coexpression analyses.


A
143
237
19
TEA-GCN
ATTED-II
TF Positive edges
Z (Co-exp.)
B
TEA-GCN
ATTED-II
JLO
BUM1
KNAT1
WRKY25
HSF4
HSFB2A
RHL41
WRKY6
TRX5
LOX3
OBP2
CYP83B1
MAM1
CYP79B2
CYP79B3
ANAC019
VSP1
ERD1
GSTF6
LOX2
COR47
ABI5
AT5G66400
ABI3
OEP16-2
SDH2-3
GEA6
EM1
FUS3
AT1G62290
AT2S3
Z (Co-exp.)
TF
Non-TF
C
Transcription factor co-expression neighbourhood
Genes ranked by co-exp.
Highest
Lowest
ES
Calculate Enrichment Score (ES)
GO-BP Terms
Hyperosmotic response
Heat acclimation
Hyperosmotic salinity response
Response to salt
Cellular reponse to salt stress
Response to cold
Response to heat
Response to reactive oxygen species
Response to osmotic stess
Resposne to salt stress
p_BH = 0.05
NES
− log10 (p_BH)
D
Enrichement Score
Genes Ranked by Co-exp.

311

Figure 4. Characterization of transcription factor-target predictions in *A. thaliana* with TEA-GCN. A) Venn diagram (top) showing 673 common and unique TF positive edges that are co-expressed (***Z***(Co-exp.) ≥ 1) in *A. thaliana* TEA-GCN and ATTED-II GCN. Heatmap (bottom) shows co-expression strengths of TF positive edges in the two GCNs. Edges are sorted based on their co-expression in ATTED-II GCN. Color intensities correspond to co-expression strength (i.e., ***Z***(co-exp.)). **B)** Co-expression between selected TFs and their known targets in TEA-GCN and ATTED-II GCN. TF and non-TF genes are represented by yellow diamond-shaped and purple circular-shaped nodes, respectively. Edge thickness represents co-expression strength between TFs and their known targets. Grey chevrons indicate the direction of regulation between TFs and their targets. **C)** Left: Schematic of enrichment score calculation from co-expression neighbourhood. Right: Dot plots comparing the enrichment statistics of selected GO terms in the neighborhood of *WRKY25* in TEA-GCN and in ATTED-II GCN. GO terms were selected based on the known functions of *WRKY25*. Dot sizes correspond to normalized enrichment scores (NES), while inverted triangles represent GO terms with negative NES scores. For GO terms with positive NES scores, the x-axis indicates the statistical significance in the form of Benjamini-Hochberg-corrected p-values (p_BH). Red dotted lines denote the statistical significance threshold (p_BH ≥ 0.05). **D)** Running-sum profiles for enrichment scores and annotated gene rankings of selected GO terms between co-expression neighborhoods of *WRKY25* in TEA-GCN and ATTED-II GCN. In D and E, the colors of the dots, running sum profiles, and annotated gene rankings correspond to the different selected GO terms.

TEA-GCN outperforms the state-of-the-art despite using a substantially smaller dataset

Given that there may be little publicly available RNA-seq data for some species, we investigated TEA-GCN performance on smaller datasets. To this end, we generated GCNs for downsampled datasets consisting of 5,000 samples and 500 samples that were randomly selected (without replacement) from the *A. thaliana* public transcriptome dataset. While the overall performance of the *A. thaliana* TEA-GCNs markedly decreases as the datasets become smaller, the *A. thaliana* TEA-GCN at *n* = 500 is still substantially more performant (AUROC-HM: 0.649, AUPRC-HM: 0.677; Figure 5A; Table S5) than the non-ensemble PCC GCN (AUROC-HM: 0.602, AUPRC-HM: 0.622; Figure 2A and B; Table S2) generated from the complete *A. thaliana* public transcriptome dataset (*n* = 57,759). Crucially, *A. thaliana* TEA-GCN at *n* = 5,000 (AUROC-HM: 0.669, AUPRC-HM: 0.699; Figure 5A; Table S5) can still outperform ATTED-II GCN (AUROC-HM: 0.659, AUPRC-HM: 0.690; Figure 5A; Table S2), despite using a downsampled dataset that is more than one order of magnitude smaller than the complete dataset. This outperformance is made more impressive considering that ATTED-II GCN is generated via the “subbaging” method using 14,741 curated, sample-balanced, and batch-corrected RNA-seq samples (according to https://atted.jp/download/), as opposed to *A. thaliana* TEA-GCN at *n* = 5,000, which uses 5,000 non-batch-corrected RNA-seq samples that have been randomly selected from the public transcriptomic dataset. Thus, the performance of TEA-GCN is not only translatable to smaller datasets but can likely outperform the “subbaging” methods, which require time-consuming and complex annotation and balancing of samples.

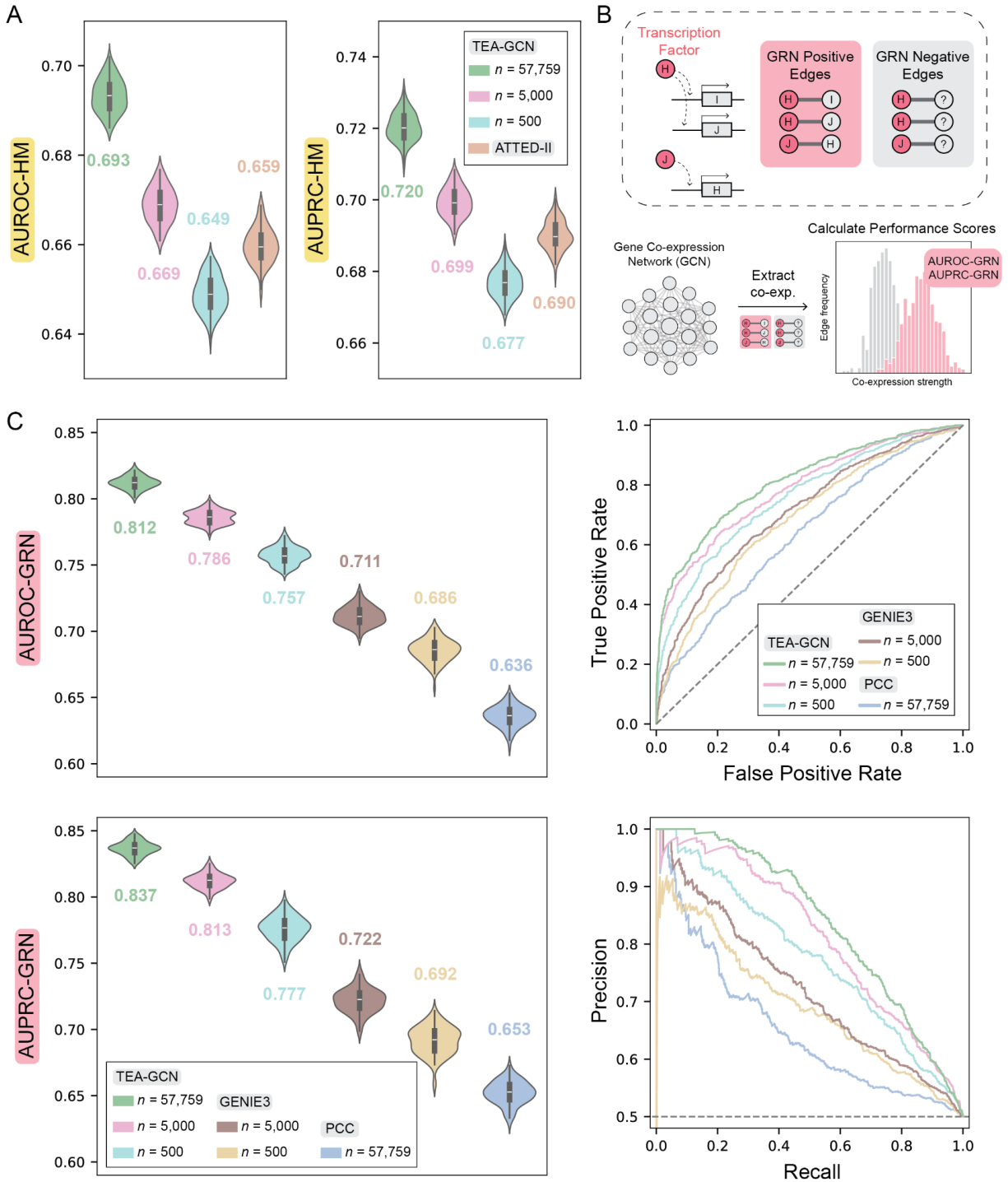


Figure 5. Scalability of TEA-GCN to smaller datasets and its GRN inference performance compared to GENIE3. A) Distribution of AUROC-HM performance scores and AUPRC-HM performance scores between ATTED-II GCNs and *A. thaliana* TEA-GCN constructed from the complete transcriptomic dataset (n = 57,759) and randomly downsampled datasets (n = 5,000 and n = 500), that were sampled 100 times. **B)** AUROC-GRN and AUPRC-GRN calculation relies on GRN positive edges and GRN negative edges. GRN positive edges connect transcription factors to the targets they activate. GRN positive edges are directional, differing from TF positive edges; thus, transcription factors targeting each other are connected by two separate positive edges. GRN negative edges are generated by randomizing genes connected to transcription factors in GRN positive edges. **C)** Distribution of AUROC-GRN and AUPRC-GRN scores of *A. thaliana* TEA-GCNs and GENIE3-GRNs constructed from the complete and downsampled transcriptomic

datasets. Non-ensemble PCC GCN is included as a comparison. Receiver operating characteristic curves (ROCs) and Precision-recall curves (PRCs) correspond to median performance scores for each GCN.

TEA-GCN outperforms state-of-the-art in GRN inference

As demonstrated earlier, the TEA-GCN framework excels in revealing gene regulatory networks (Figures 2C, 2D, and 4), which prompted us to compare its GRN-inference capabilities to GENIE3[24], a well-established machine learning-based algorithm dedicated to inferring GRN from transcriptomic data. Comparison of TEA-GCN and GENIE3 revealed that TEA-GCN outperforms GENIE3 in identifying directional transregulatory relationships (GRN positive edges, Figure 5B, Table S6) even when a smaller dataset comprising < 10% of samples is used (Figure 5C; Table S7). TEA-GCN at n = 500 (AUROC-GRN: 0.757, AUPRC-GRN: 0.777; Figure 5C; Table S7) outperforms the GENIE3 GRN at n = 5,000 (AUROC-GRN: 0.711, AUPRC-GRN: 0.722; Figure 5C; Table S7) in GRN inference despite a ten-fold difference in dataset size. *A. thaliana* TEA-GCN generated using the complete dataset (n = 57,759) has the best GRN-inference ability (AUROC-GRN: 0.812, AUPRC-GRN: 0.837; Figure 5C; Table S7). On the other hand, the non-ensemble PCC GCN generated using the complete dataset (n = 57,759) performed the worst (AUROC-GRN: 0.636, AUPRC-GRN: 0.653; Figure 5C; Table S7). Thus, TEA-GCNs can outperform dedicated GRN algorithms, such as GENIE3, in GRN inference.

Experimental contexts underpinning gene co-expression can be explained in TEA-GCN

Despite their utility in identifying functional associations between genes, current network-based methods cannot explain which biological or experimental conditions (e.g., organ, tissue, treatment) underpin the observed co-expression between genes. Indeed, machine-learning GRN methods such as GENIE3[54] and GRNBoost2[55], are often considered “black boxes” due to the difficulty in understanding how the network edges are determined.

To investigate whether TEA-GCN provides more explainability for gene-gene relationships than these machine-learning methods, we analyzed the datasets underpinning the condition/tissue-specific co-expression analysis (Figure 6). To this end, we developed a Natural Language Processing (NLP) method that analyzes the metadata for each of the data partitions identified by k-means (Figure 6A), identifies statistically significant enriched keywords in each partition (overrepresented lemmas; see Methods; Figure 6A, Table S8), and heuristically

390 annotates ensemble co-expression edges with co-expression contexts based on their underlying
391 co-expression profile across partitions (Figure 6B; Table S9). To test our approach, we applied
392 this method to GRN-positive edges that exhibited strong co-expression in the *A. thaliana* TEA-
393 GCN (i.e., ***Z***(Co-exp.) ≥ 1.5). The majority of identified edges (293 out of 400 edges; 73.25%;
394 Table S10) could be annotated with at least one lemma, and these annotations largely correspond
395 with the known functions of the gene pairs (Figure 6C; Table S10). To exemplify, we selected two
396 well-characterized transcription factors, ABSCISIC ACID INSENSITIVE 5 (ABI5) and MALE
397 STERILE 188 (MS188). Five positive edges of *ABI5* (connected to *GEA6*, *EM1*, *SDH2-3,* and
398 *AT5G66400)* were annotated with abscisic acid (ABA), a phytohormone that mediates important
399 physiological processes such as seed dormancy maintenance and abiotic stress response[56–58]
400 (Figure 6C; Table S11). This is consistent with the prominent role of ABI5 in regulating ABA
401 signaling[45,59,60]. Interestingly, *ABI5* edges connecting *AT5G66400*, *EM1*, and *GEA6* were
402 annotated with experimental contexts reflecting low-light/dark growth conditions (i.e., "Shade",
403 "Etiolate", "Darkness"; Figure 6C; Table S11), which represents novel information that can be
404 used to further dissect the regulatory wirings underpinning ABA signaling of ABI5. The revealed
405 darkness-related experimental contexts are supported by the recent identification ofCOP1 E3
406 ligase[61] and phytochrome-interacting factors[62] as darkness-specific modulators of ABI5 activity.
407 In another example, all three edges of *MS188* (connected to *FAR2*, *MYB35*, and *MS1*)
408 were annotated with experimental contexts related to floral tissues (i.e. "Flower", "Flowering",
409 "Inflorescence"), which is congruent with MS188's known function in anther developmental
410 processes (e.g., tapetum development, pollen exine formation, and pollen callose development;
411 Figure 6C; Table S11)[63–65]. Thus, TEA-GCN, when combined with NLP approaches, produces
412 interpretable edges that allow the elucidation of the experimental context, further boosting gene
413 function prediction approaches.

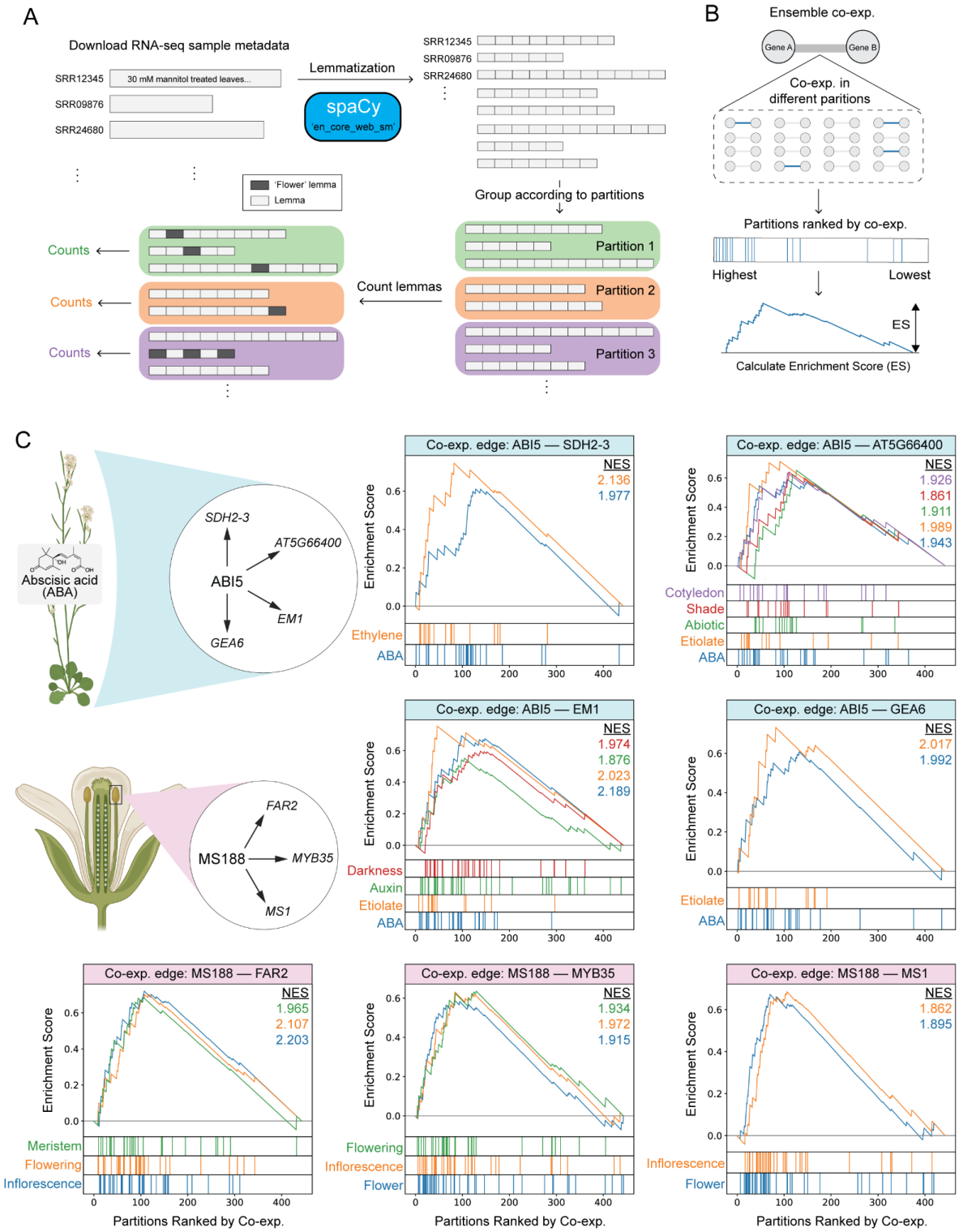


Figure 6. Annotating co-expression edges of *A. thaliana* TEA-GCN with experimental contexts. A) The metadata of RNA-seq samples is broken down into lemmas. Lemmas of RNA-seq samples belonging to the same dataset partition are then combined. For each partition, occurrences of different lemmas are counted, and enriched lemmas were determined by comparing the observed lemma occurrences to permuted occurrences when the partition assignments of RNA-seq samples are randomly shuffled. Dataset partitions are annotated by overrepresented lemmas denoting the experimental conditions or tissues from which the samples in the

dataset partitions were predominantly generated. **B)** Co-expression edges of TEA-GCN can be annotated with lemmas enriched with an approach similar to Gene-set Enrichment Analysis (GSEA). The ensemble co-expression strength of a given TEA-GCN edge is first deconvoluted into co-expression strengths obtained from the different dataset partitions, and these strengths are then used to rank the dataset partitions. This approach facilitates the calculation of enrichment scores, as seen in GSEA. **C**) Lemma enrichment analysis of edges connecting *ABSCISIC ACID INSENSITIVE 5* (*ABI5*) and *MALE STERILE 188* (*MS188*) transcription factors to their known targets. Running-sum profiles for enrichment scores and annotated partition rankings of the co-expression context lemmas detected for each of the presented edges are shown. The line colors of the running sum profiles correspond to the colors of experimental context lemmas.

TEA-GCNs Networks Are Highly Conserved Across Species

The immense variability of publicly available RNA-seq data amongst different species may result in poor agreement (conservation) of GCNs, even between closely-related species[1,2], resulting in their limited utility in cross-species studies. To determine if TEA-GCN networks suffer from the same problem, , we analyzed TEA-GCNs from ten diverse Angiosperm species, including *A. thaliana* (Figure 7A), and evaluated cross-species comparability. Our selected panel includes both major crops—*Brassica rapa ssp. pekinesis* (Napa cabbage), *Vitis vinifera* (Grape), *Solanum lycopersicum* (Tomato), *Oryza sativa* (Japanese rice), and *Zea mays* (Maize)—and key evolutionary models such as *Medicago truncatula* (Barrel clover) and *Populus trichocarpa* (Black cottonwood). To enable GCN comparisons between closely related species, we also included *Arabidopsis lyrata* (*A. lyrata*) and *Arabidopsis halleri (A. halleri)*, which belong to the same genus as *A. thaliana*.

To benchmark GCN performance, we defined a set of Positive Met edges for each species (except *A. thaliana*) based on PlantCyc[29] enzyme annotations, analogous to the Positive Met edges used for *A. thaliana* (Figure 1B). TEA-GCNs consistently outperform conventional non-ensemble PCC-GCNs across all tested species (Figure 7B). Notably, the PCC-GCNs for *A. lyrata* and *A. halleri* perform poorly, approaching random classification where AUROC and AUPRC are 0.5, while the TEA-GCNs for the same species perform substantially better. The TEA-GCNs, by contrast, show more consistent and robust performance across all three Arabidopsis species, underscoring their reliability even in data-scarce settings (Figure 7B).

To systematically assess cross-species comparability, we developed two complementary conservation metrics based on orthologous gene pairs (Figure 7C): (1) Co-expression correlation, which measures the rank-based similarity in edge weights between GCNs via Spearman's rank Correlation Coefficient (SCC), and (2) Co-expression overlap, which quantifies the agreement in highly co-expressed edges, based on ***Z***(Co-exp. str.) > 1.5, using the Jaccard Index. Applying these metrics to all pairwise combinations of the 10 species (45 comparisons total), we find that

TEA-GCNs exhibit markedly higher conservation than PCC-GCNs across both metrics (Figure 7D). This suggests that TEA-GCNs not only recover biologically relevant relationships more effectively within species but also yield networks that are more consistent across species boundaries—an essential feature for evolutionary and comparative analyses. To exemplify, the conservation metrics between *A. halleri* (28,722 genes) and *A. thaliana* (27,562 genes)—based on nearly 200 million orthologous edges from 19,989 identified 1-to-1 orthologs—dramatically improved under TEA-GCNs. Specifically, the Co-expression correlation increased from 0.415 (PCC-GCN) to 0.500 (TEA-GCN), and the Co-expression overlap doubled from 0.178 to 0.354 (Figure 7D). In other words, the number of conserved, highly co-expressed orthologous edges between these two species doubled when using TEA-GCNs. These results collectively demonstrate that TEA-GCNs yield more biologically meaningful and evolutionarily conserved gene co-expression, positioning them as a powerful framework for cross-species transcriptomic analyses.

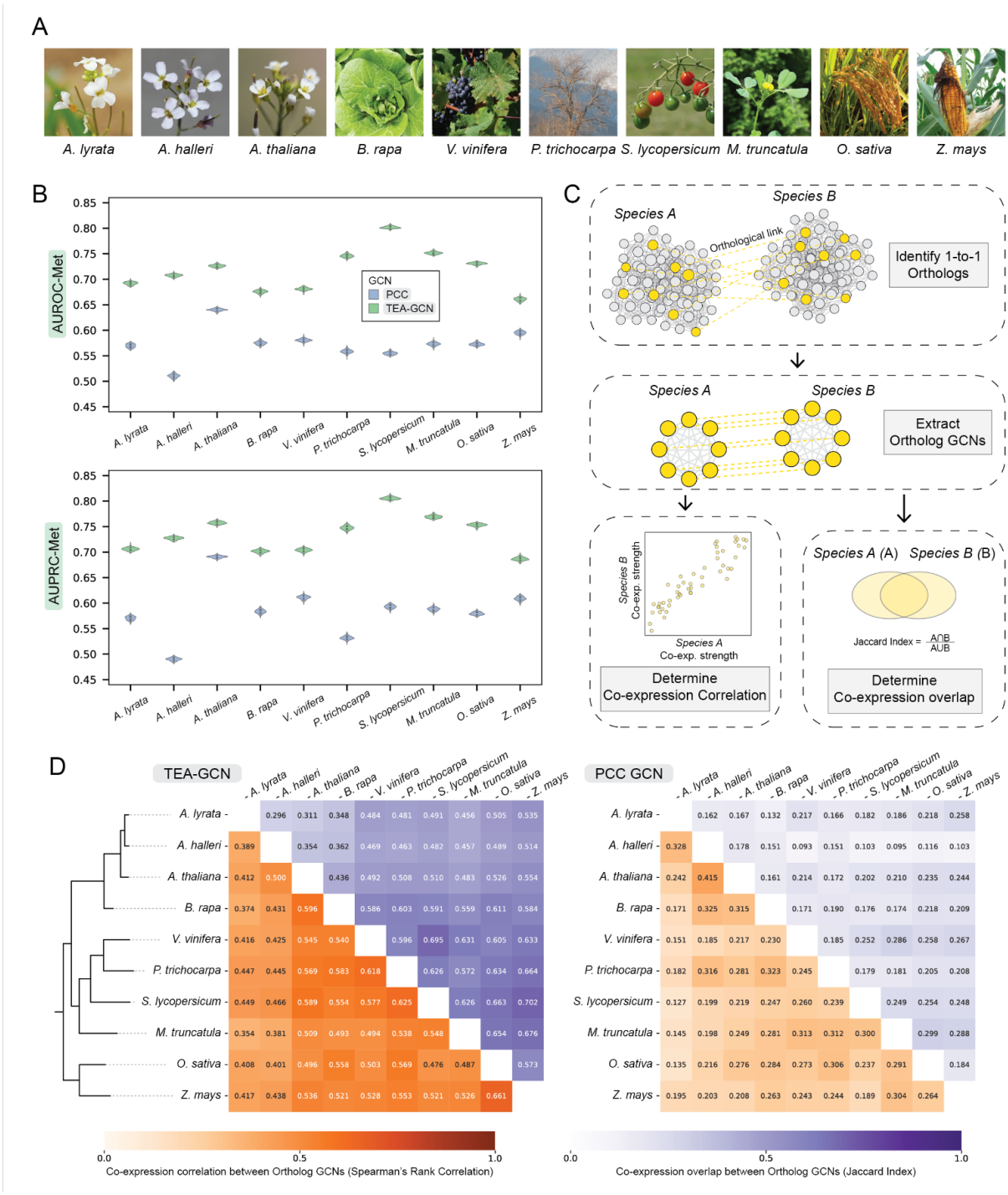


Figure 7. TEA-GCNs demonstrate superior performance and conservation across ten Angiosperm plant species compared to non-ensemble PCC GCNs. A) Images of the ten Angiosperm Plant Species. All images are sourced from the public domain. Creative Commons licenses and creator attributions (Table S25). **B)** AUROC-Met and AUPRC-Met performance scores of PCC-GCNs (blue) and TEA-GCNs (green) of various angiosperm plant species. **C)** Schematic overview of the conservation analysis pipeline. Subgraphs consisting of 1-to-1 ortholog pairs (Ortholog GCNs) are extracted from species GCNs. Two metrics are used to assess GCN similarity: (1) Co-expression correlation (Spearman's Rank Correlation Coefficient of edge weights), and (2) Co-expression

overlap (Jaccard Index of edges exceeding a co-expression threshold). Orthologs are defined using OrthoFinder2. **D)** Pairwise conservation of TEA-GCNs (left) and PCC GCNs (right) across all species, shown as heatmaps. The top triangle of each heatmap shows Co-expression correlation, while the bottom triangle shows Co-expression overlap. Dendrograms reflect phylogenetic relationships estimated by OrthoFinder2.

Discussion

Gene co-expression networks (GCNs) generated from transcriptomic datasets can elucidate the intricate co-regulatory and co-functional relationships between genes, thus enabling gene function prediction. Consequently, GCNs have proven invaluable as a hypothesis generation tool in non-model species, where gene functions are hard to predict based on sequence similarity to characterized homologs in well-studied models (i.e., members of large gene families, orphan and lineage-specific genes/proteins)[66–70]. Due to the increased affordability of next-generation sequencing (NGS), the number of species with transcriptomic data is growing rapidly, which presents an unprecedented opportunity to unravel the evolutionary dynamics of gene function across diverse organisms via the construction of GCNs. However, current GCN construction methodologies[3,26,71,725] were developed for use on small in-house sequenced datasets or manually curated public RNA-seq samples and, therefore, are not well equipped to handle (nor to exploit) the various complexities of public transcriptome data (e.g., high heterogeneity, sample imbalance, batch effects, poor sample quality). To address this unmet need, we developed TEA-GCN, an ensemble methodology centered around determining gene co-expression based on local co-expression observed in different dataset partitions (Figure 1).

TEA-GCN is superior to the widely-used method of calculating PCC across the whole dataset (PCC GCN)[26,71,73,74], and to the state-of-the-art “Subbagging” method from the ATTED-II and COXPRESdb databases[15,18] in identifying transregulatory relationships as well as gene co-involvement in enzymatic pathways and biological processes, across three model organisms of vastly different organizational/biological complexities and evolutionary lineages (Figures 2 and S3). In addition, TEA-GCNs possess enhanced across-species comparability and stability compared to PCC GCNs as seen from their higher conservation of network features across ten Angiosperm species (Figure 7), which can better facilitate kingdom-wide evolutionary studies across GCNs from different species. Evident from the characterization of co-expression edges, TEA-GCN excels in capturing co-expression relationships that manifest in certain conditions and/or in specific tissues, which enables the discovery of complex regulatory mechanisms that can be elusive (Figures 3 and 4). In addition, known functions of seven condition/tissue-specific

508 transcription factors (i.e., *WRKY25*, *OBP2*, *ANAC091*, *JLO*, *FUS3*, *WRKY6*, and *ABI5*) can be
509 predicted in their co-expression neighborhoods in TEA-GCN but not in the ensemble GCN from
510 ATTED-II (Figures 4 and S4), exemplifying TEA-GCN's enhanced ability to uncover gene function
511 elusive to traditional co-expression. Furthermore, the targets of TOR kinase, identified via protein-
512 proximity labeling in *A. thaliana*, in another study[75] are strongly corroborated by our *A. thaliana*
513 TEA-GCN. Inspired by the growing field of Explainable Artificial intelligence[76] (XAI) toward
514 explaining the inner workings of complex machine-learning models, TEA-GCN can be made to
515 explain experimental contexts underpinning gene co-expression via NLP. This was possible by
516 capitalizing on our method's ability to capture condition/tissue-specific co-expression. Thus, TEA-
517 GCN can not only outperform more complex state-of-the-art methods but also, for the first time,
518 report biologically significant information as experimental contexts to contextualize gene co-
519 expression.

520 TEA-GCN is designed to operate on complex and heterogeneous public transcriptomic
521 datasets where traditional co-expression inference using PCC is suboptimal. Similar complexity
522 is observed in single-cell RNA-sequencing (scRNA-seq) datasets, which comprise tens of
523 thousands of transcriptomes from diverse cell types. In these datasets, gene expression
524 measurements are often sparse due to limited sequencing depth per cell, which hampers accurate
525 co-expression inference using PCC[77]. Therefore, there is significant potential for extending TEA-
526 GCN to infer gene co-expression networks (GCNs) from scRNA-seq data, highlighting a key
527 direction for future development. Although the ensemble methodology of TEA-GCN vastly
528 improves the quality of GCNs, various aspects of TEA-GCN can still be further optimized. For
529 example, the K-means clustering algorithm was chosen to identify biologically-relevant partitions
530 from datasets for TEA-GCN as the sizes of the yielded partitions are more uniform as compared
531 to other popular clustering methods such as the Density-based spatial clustering of applications
532 with noise (DBSCAN) algorithm[78] or the various algorithm variants for Hierarchical Agglomerative
533 Clustering[79]. A systematic evaluation of the different clustering methods can help identify ways to
534 generate partitions that can lead to better downstream TEA-GCN performance. In addition, a
535 straightforward method was employed in the study to annotate dataset partitions based on
536 statistically overrepresented lemmas in RNA-seq sample metadata when determining
537 experimental contexts underpinning TEA-GCN ensemble co-expression. In the face of rapid
538 advancements in generative LLMs[80,81], one can tap into LLMs to annotate dataset partitions with
539 greater accuracy and beyond single-worded lemmas, when operational costs become less
540 prohibitive. Nonetheless, the robust compatibility of TEA-GCN to work on highly variable datasets
541 without the need for sample annotations for batch correction and/or sample balancing cannot be

542 understated as high-quality TEA-GCNs can be generated for any species with publicly available
543 RNA-seq samples. This opens the door for scientists working on these species to utilize co-
544 expression to accelerate their research as well as enabling analyses to be expanded from well-
545 studied models into other species in the same evolutionary group to uncover the evolutionary
546 origins of genes and pathways. The lower barrier to generate quality GCNs using our method can
547 also facilitate the generation of multi-species co-expression networks for downstream analyses.
548 For instance, one can construct genus/clade-specific GCNs (e.g., GCN of cruciferous vegetables)
549 or even GCNs for the Human Pangenome[82]. To take this even further, novel algorithms can be
550 developed to reconstruct ancestral gene co-expression modules from a compendium of extant
551 TEA-GCNs similar to how ancestral genomes can be inferred from extant genomes[83].
552 The TEA-GCN method is available for use via https://github.com/pengkenlim/TEA-GCN
553 as an easy-to-install and well-documented standalone program for users to generate high-quality
554 GCNs from their large transcriptomic datasets. Additionally, the TEA-CGNs of ten Angiosperm
555 species are available in Plant-GCN via https://plantgcn.connectome.tools , a web resource
556 equipped with user-friendly tools for exploration and analysis (Supplementary results, Figure S6
557 to S8).

558 # Methods

559 **Generation and preprocessing of *S. cerevisiae*, *A. thaliana,* and *H. sapiens* public**
560 **transcriptomic datasets**

561 *S. cerevisiae* and *H. sapiens* mRNA sequences were downloaded from the NCBI RefSeq
562 annotations of their reference genome assemblies (GCF_000146045.2 and GCF_000001405.40).
563 Representative mRNA sequences of protein-coding genes (primary transcripts) for *S. cerevisiae*
564 and *H. sapiens* were then obtained by selecting the longest transcript isoform for every protein-
565 coding gene locus. Primary transcripts of *A. thaliana* were obtained by downloading the
566 "Araport11_cdna_20220914_representative_gene_model.gz" file from the TAIR database[84].
567 6,612 *S. cerevisiae*, 19,379 *H. sapiens*, and 27,562 *A. thaliana* primary transcripts were used as
568 pseudoalignment references for gene expression estimation of RNA-seq sample Illumina
569 sequencing reads (read data) downloaded from ENA using Kallisto's quant functionality in single-
570 end mode with default parameters[85]. Fastq files of 188,189, 66,048, and 142,828 RNA-seq
571 samples were downloaded and pseudoaligned for *S. cerevisiae, A. thaliana, and H. sapiens,*
572 respectively. Transcripts per million (TPM) values of primary transcripts of RNA-seq samples
573 estimated by Kallisto were used as gene expression abundances to construct gene expression
574 matrices for the different species. RNA-seq samples with poor read alignment were identified
575 based on their %_pseudoaligned statistics (reported by Kallisto) below a certain threshold (5%
576 for *H. sapiens*; 20% for *S. cerevisiae* and *A. thaliana*) as described[86–88]. After the removal of RNA-
577 seq samples with poor read alignment, gene expression matrices comprising of gene expression
578 abundances from 150,096 (79.8%), 57,759 (87.5%), and 95,553 (66.9%) retained RNA-seq
579 samples were designated as the public transcriptomic datasets for *S. cerevisiae, A. thaliana,* and
580 *H. sapiens,* respectively.
581

582 **Data sources for establishing positive edges**

583 Experimentally verified relationships (positive edges) for Met, TF, and GO categories were
584 established for the three model species (Tables S12 to S20). *S. cerevisiae* TF positive edges are
585 generated by downloading activation regulatory links from the YEASTRACT+ database[89]
586 (http://www.yeastract.com/) that fulfill both"Documented" and "Potential" criteria. According to the
587 authors, "Documented " regulatory links are inferred from experimental transcription factor-
588 binding evidence, while "Potential" regulatory links connect transcription factors with target genes

589 that contain corresponding consensus sequences in the upstream promoter region[89]. *A. thaliana*
590 TF positive edges are generated by downloading activation regulatory links from the PlantTFDB
591 database[90] (https://planttfdb.gao-lab.org/). *H. sapiens* TF positive edges are generated by
592 downloading activation regulatory links from the TRRUST database[91] (version 2;
593 https://www.grnpedia.org/trrust/) and then eliminating regulatory links that do not connect to any
594 known *H. sapiens* transcription factors. *H. sapiens* transcription factors are obtained from the
595 HumanTFs database[92] (https://humantfs.ccbr.utoronto.ca/). GO annotations for S. *cerevisiae*, *A.*
596 *thaliana,* and *H. sapiens* were downloaded from EBI Gene Ontology Annotation (GOA)
597 Database[93], The Arabidopsis Information Resource[84] (TAIR), and *Saccharomyces* Genome
598 Database[94] (SGD), respectively. Pathway annotations for S. *cerevisiae*, *A. thaliana,* and *H.*
599 *sapiens* were webscraped from BioCyc[95], PlantCyc[29], and HumanCyc[96], respectively. Since there
600 is a tendency for co-linearity between the estimated expression abundances of sequence-similar
601 homologs due to how multi-mapping reads are accounted for in pseudoalignment[85], edges that
602 connect genes annotated with the same molecular function were removed from GO positive
603 edges to prevent these edges from inflating performance scores (Figure 1B; blue). Similarly, we
604 removed edges that connect genes coding for enzymes that catalyze the same reaction steps
605 from Met positive edges.
606 Similar to *A. thaliana*, the Positive Met edges of other Angiosperm species in this study
607 were mined from PlantCyc[29]. Since experimentally validated enzymes annotations are scarce or
608 non-existent for the Angiosperm plant species (excluding *A. thaliana*), the definition of these
609 Positive Met edges were broadened for said species to include predicted enzyme-gene
610 associations while excluding edges involving enzymes predicted to catalyze highly promiscuous
611 reactions (i.e., reactions with > 5 predicted enzymes) to control false positive edges (Table 21).
612 These edges were excluded on the assumption that enzymes annotated to catalyze promiscuous
613 reactions are more likely to be inaccurate.
614
615 **Calculation of AUROC and AUPRC Scores to evaluate GCN performance**
616 Using the co-expression strengths extracted for different sets of positive edges and negative
617 edges (representing genes with no functional relationship, obtained from sampling random gene
618 pairs), the Area-under-curve performance scores of the receiver operator characteristic curve
619 (AUROC) and the precision-recall curve (AUPRC) were calculated for each biological aspect
620 (e.g. ,AUROC-Met and AUROC-Met; Figure 1C). We aggregated the scores with harmonic mean
621 to infer a GCN's overall performance (i.e., AUROC-HM and AUROC-HM; Figure 1C). To account

for the random effect on performance scores due to the random selection of negative edges, median scores from 100 evaluation iterations are reported.

Standardizing co-expression strengths within GCNs

To compare the co-expression strengths of edges connecting the same gene pair across different GCNs, the co-expression strengths of edges were standardized before comparison. Because different CGNs do not necessarily contain the same genes and edges, the co-expression strengths of each GCN were standardized based on the mean and standard deviation of the co-expression strengths from a subset of common edges shared between the GCNs. The mathematical description of said standardization is detailed below.

Let $X = [x_1, x_2, x_3 \ldots, x_N]$ be a vector of co-expression strengths of common edges in one of the GCNs to be compared. The mean co-expression strength (μ) and standard deviation (σ) of X can be calculated as follows:

$$\mu = \frac{\sum_i x_i}{N}$$

$$\sigma = \sqrt{\frac{\sum(X - \mu)}{N}}$$

The co-expression strength of any given edge in the GCN (x_i) can then be standardized as follows:

$$Z(x_i) = \frac{x_i - \mu}{\sigma}$$

For interpretability, ***Z***(co-exp.) is used to refer to standardized co-expression strengths throughout the text.

GENIE3 GRN inference and evaluating GRNs

The Python implementation of GENIE3[54] was used to infer GRNs used in this study with default settings (i.e., ntrees = 1000, tree-method = Random Forest, K = square root method). A list of 324 *A. thaliana* transcription factors (Table S6) was obtained from PlantTFDB[90] and used as regulators for GENIE3. GRN-inference methods like GENIE3 calculate the propensities of genes as regulatory targets for transcription factors defined by the user. Since GRNs are directional (e.g., TF1 → TF2) and the aim is to predict TF-target gene pairs (rather than any gene-gene

651 pairs), we defined a new set of positive directional edges from experimentally verified GRNs for
652 *A. thaliana* (Figure 5B; Table S6). Negative edges used to calculate AUROC-GRN and AUPRC-
653 GRN were defined by only randomizing target genes connected to edges in the GRN positive
654 edge set (Figure 5B) rather than edges simply connecting two random genes. AUROC-GRN and
655 AUPRC-GRN were calculated 100 times from propensities (for GENIE3) and co-expression
656 strengths (for TEA-GCN) of edges (Figure 5B).

657 **Downloading of *A. thaliana* RNA-seq metadata**

658 The metadata of RNA-seq samples was downloaded from ENA via the ENA Browser API using
659 "sample_description", "dev_stage", "study_title", "sample_title", "tissue_lib", and "tissue_type" as
660 search fields (Tables S22 to 23).

661 **Identifying overrepresented lemmas in *A. thaliana* dataset partitions**

662 The "en_core_web_sm" NLP model (version 3.7.1) from the SpaCy (version 3.7.1) Python
663 package was used to lemmatize metadata of *A. thaliana* RNA-seq samples. Before lemmatization,
664 the metadata of RNA-seq samples was preprocessed. First, metadata returned from all search
665 fields of every RNA-seq sample were concatenated. Next, stop words (e.g., about, quite, the)
666 were detected from the concatenated metadata using the "token.is_stop" attribute of word tokens
667 generated by the "en_core_web_sm" model. Digits and punctuation were also removed from the
668 concatenated metadata.

669 To obtain overrepresented lemmas from the different dataset partitions, metadata lemmas
670 of RNA-seq samples belonging to the same partition were grouped and combined to represent
671 the lemmas of each dataset partition. The occurrence of lemmas in each dataset partition was
672 then counted and stored as observed lemma occurrences. Permuted lemma occurrences were
673 similarly obtained by repeating the counting process on dataset partition lemmas combined from
674 randomly shuffled sample-partition assignments. A null hypothesis p-value indicating the
675 overrepresentation significance of a particular lemma in a dataset partition can be mathematically
676 defined below.

677 Let $S = [s_1, s_2, s_3, \ldots s_n]$ be a vector of permuted occurrences of a lemma in a dataset
678 partition from n random shufflings of sample-partition assignments, and o be the observed
679 occurrence. The p-value (P) can be calculated as follows:

680 $$P = \frac{1 + \sum_{i=1}^{n} I(s_i \geq o)}{n}$$

Where I is an indicator function that takes a value of 1 or 0 if the statement is true or false, respectively.

P-values determined for every lemma in each dataset partition were then subjected to multiple-testing correction using the Benjamini-Hochberg method[97], yielding Benjamini-Hochberg corrected p-values (p_BH). Based on 1000 permutations, 2,333 overrepresented lemmas (p_BH ≤ 0.05, o > 20) were identified for the *A. thaliana* dataset partitions. After manual selection to remove lemmas that are unlikely to contribute to biological insight (e.g., "arabidopsis", "bottle", "clean", "device", etc.), 276 lemma labels remain (Table S8).

Enrichment-based analyses

GSEA of co-expression neighborhoods of selected *A. thaliana* transcription factors was implemented by the GSEApy python package[98] (version 1.1.2). For each transcription factor, genes were ranked based on co-expression strength. The ranks were centered, and gene sets were constructed using *A. thaliana* GO annotations (biological process category). Together, the centered ranks and constructed gene sets were used as input for the prerank subcommand of GSEApy (weight = 1). Annotating edges between co-expressed genes with experimental context lemmas also uses the prerank sub-command of GSEApy in the same way. However, centered partition rankings determined from ranking dataset partitions based on co-expression strengths after Coefficient Aggregation were used in place of gene rankings. Instead of gene sets, sets of dataset partitions for which lemma labels were assigned were used.

Calculating GCN conservation between species

For GCNs of any two species, subgraphs that contain only edges connecting 1-to-1 orthologs shared between the two species can be extracted as Ortholog GCNs (Figure 7C). Co-expression correlations between two GCNs were calculated based on Spearman's rank Correlation Coefficient[99] with edge co-expression strengths between their ortholog GCNs as two separate sets of observations. On the other hand, Co-expression Overlap can be calculated based on the Jaccard Index. The mathematical description for calculating Co-expression Overlap is detailed below:

$$Co\text{-}expression\ Overlap = \frac{|A \cap B|}{|A \cup B|}$$

710 Where A and B are sets of co-expressed edges in the Ortholog GCNs of the two species
711 with a co-expression threshold of ***Z***(Co-exp.) ≥ 1.5. Orthology between Angiosperm genes (Figure
712 S5) was inferred from OrthoFinder2[100] based on their primary transcript sequences (Supplemental
713 methods).

714 Data and code availability

715 TEA-GCN is available at https://github.com/pengkenlim/TEA-GCN. *S. cerevisiae, A. thaliana*, and
716 *H. sapiens* TEA-GCNs generated by this study can be found at FigShare DOIs
717 10.6084/m9.figshare.26347018, 10.6084/m9.figshare.26347078, and
718 10.6084/m9.figshare.26347297, respectively. Transcriptomic datasets in the form of gene
719 expression matrices used in this study can be found under 10.6084/m9.figshare.26346916,
720 10.6084/m9.figshare.26346964, and 10.6084/m9.figshare.263470105, respectively.
721 Supplemental tables referenced in this study are available at 10.6084/m9.figshare.29071076.
722 TEA-GCNs for 10 Angiosperms can be accessed via the Plant-GCN database at
723 https://plantgcn.connectome.tools/ .

724 # References

725 1. Usadel, B. *et al.* Co-expression tools for plant biology: opportunities for hypothesis
726 generation and caveats. *Plant Cell Environ.* **32**, 1633–1651 (2009).
727 2. Lim, P. K., Zheng, X., Goh, J. C. & Mutwil, M. Exploiting plant transcriptomic databases:
728 Resources, tools, and approaches. *Plant Commun.* **3**, 100323 (2022).
729 3. Rao, X. & Dixon, R. A. Co-expression networks for plant biology: why and how. *Acta*
730 *Biochim. Biophys. Sin.* **51**, 981–988 (2019).
731 4. Aoki, K., Ogata, Y. & Shibata, D. Approaches for extracting practical information from gene
732 co-expression networks in plant biology. *Plant Cell Physiol.* **48**, 381–390 (2007).
733 5. Rhee, S. Y. & Mutwil, M. Towards revealing the functions of all genes in plants. *Trends*
734 *Plant Sci.* **19**, 212–221 (2014).
735 6. Julca, I., Tan, Q. W. & Mutwil, M. Toward kingdom-wide analyses of gene expression.
736 *Trends Plant Sci.* **28**, 235–249 (2023).
737 7. Burgin, J. *et al.* The European Nucleotide Archive in 2022. *Nucleic Acids Res.* **51**, D121–
738 D125 (2023).
739 8. Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species.
740 *Nucleic Acids Res.* **50**, D129–D140 (2022).
741 9. Ruprecht, C., Vaid, N., Proost, S., Persson, S. & Mutwil, M. Beyond Genomics: Studying
742 Evolution with Gene Coexpression Networks. *Trends Plant Sci.* **22**, 298–307 (2017).
743 10. Ovens, K., Eames, B. F. & McQuillan, I. Comparative Analyses of Gene Co-expression
744 Networks: Implementations and Applications in the Study of Evolution. *Front. Genet.* **12**,
745 695399 (2021).
746 11. Vandenbon, A. Evaluation of critical data processing steps for reliable prediction of gene co-
747 expression from large collections of RNA-seq data. *PloS One* **17**, e0263344 (2022).

748 12. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network
749 construction and analysis: safety in numbers. *Bioinforma. Oxf. Engl.* **31**, 2123–2130 (2015).
750 13. Gupta, C. & Pereira, A. Recent advances in gene function prediction using context-specific
751 coexpression networks in plants. *F1000Research* **8**, F1000 Faculty Rev-153 (2019).
752 14. Pearce, S., Ferguson, A., King, J. & Wilson, Z. A. FlowerNet: a gene expression correlation
753 network for anther and pollen development. *Plant Physiol.* **167**, 1717–1730 (2015).
754 15. Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an animal
755 gene coexpression database navigating from a global view to detailed investigations.
756 *Nucleic Acids Res.* **51**, D80–D87 (2023).
757 16. Lin, H., Yu, J., Pearce, S. P., Zhang, D. & Wilson, Z. A. RiceAntherNet: a gene co-
758 expression network for identifying anther and pollen development genes. *Plant J. Cell Mol.*
759 *Biol.* **92**, 1076–1091 (2017).
760 17. Lancour, D. *et al.* Analysis of brain region-specific co-expression networks reveals clustering
761 of established and novel genes associated with Alzheimer disease. *Alzheimers Res. Ther.*
762 **12**, 103 (2020).
763 18. Obayashi, T., Hibara, H., Kagaya, Y., Aoki, Y. & Kinoshita, K. ATTED-II v11: A Plant Gene
764 Coexpression Database Using a Sample Balancing Technique by Subagging of Principal
765 Components. *Plant Cell Physiol.* **63**, 869–881 (2022).
766 19. Liesecke, F. *et al.* Improved gene co-expression network quality through expression dataset
767 down-sampling and network aggregation. *Sci. Rep.* **9**, 14431 (2019).
768 20. Wong, D. C. J. Network aggregation improves gene function prediction of grapevine gene
769 co-expression networks. *Plant Mol. Biol.* **103**, 425–441 (2020).
770 21. Arshad, Z. & McDonald, J. F. A computational approach to generate highly conserved gene
771 co-expression networks with RNA-seq data. *STAR Protoc.* **3**, 101432 (2022).
772 22. Liesecke, F. *et al.* Ranking genome-wide correlation measurements improves microarray
773 and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* **8**, 10885

774 (2018).
775 23. Hou, J. *et al.* Distance correlation application to gene co-expression network analysis. *BMC*
776 *Bioinformatics* **23**, 81 (2022).
777 24. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual
778 information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328 (2012).
779 25. Hardin, J., Mitani, A., Hicks, L. & VanKoten, B. A robust measure of correlation between two
780 genes on a microarray. *BMC Bioinformatics* **8**, 220 (2007).
781 26. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
782 analysis. *BMC Bioinformatics* **9**, 559 (2008).
783 27. Aravind, L., Anantharaman, V. & Venancio, T. M. Apprehending multicellularity: regulatory
784 networks, genomics, and evolution. *Birth Defects Res. Part C Embryo Today Rev.* **87**, 143–
785 164 (2009).
786 28. Anantharaman, V., Iyer, L. M. & Aravind, L. Comparative Genomics of Protists: New Insights
787 into the Evolution of Eukaryotic Signal Transduction and Gene Regulation*. *Annu. Rev.*
788 *Microbiol.* **61**, 453–475 (2007).
789 29. Hawkins, C. *et al.* Plant Metabolic Network 15: A resource of genome-wide metabolism
790 databases for 126 plants and algae. *J. Integr. Plant Biol.* **63**, 1888–1905 (2021).
791 30. Delli-Ponti, R., Shivhare, D. & Mutwil, M. Using Gene Expression to Study Specialized
792 Metabolism-A Practical Guide. *Front. Plant Sci.* **11**, 625035 (2020).
793 31. Brockington, S. F. *et al.* Lineage-specific gene radiations underlie the evolution of novel
794 betalain pigmentation in Caryophyllales. *New Phytol.* **207**, 1170–1180 (2015).
795 32. Wink, M. Evolution of secondary metabolites from an ecological and molecular phylogenetic
796 perspective. *Phytochemistry* **64**, 3–19 (2003).
797 33. Lim, P. K., Julca, I. & Mutwil, M. Redesigning plant specialized metabolism with supervised
798 machine learning using publicly available reactome data. *Comput. Struct. Biotechnol. J.* **21**,
799 1639–1650 (2023).

800 34. Colinas, M. & Goossens, A. Combinatorial Transcriptional Control of Plant Specialized
801 Metabolism. *Trends Plant Sci.* **23**, 324–336 (2018).
802 35. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends Plant*
803 *Sci.* **24**, 83–98 (2019).
804 36. Voß, U., Bishopp, A., Farcot, E. & Bennett, M. J. Modelling hormonal response and
805 development. *Trends Plant Sci.* **19**, 311–319 (2014).
806 37. Zhang, Y., Berman, A. & Shani, E. Plant Hormone Transport and Localization: Signaling
807 Molecules on the Move. *Annu. Rev. Plant Biol.* **74**, 453–479 (2023).
808 38. Verma, V., Ravindran, P. & Kumar, P. P. Plant hormone-mediated regulation of stress
809 responses. *BMC Plant Biol.* **16**, 86 (2016).
810 39. Prieto, M. A., López, C. J. & Simal-Gandara, J. Glucosinolates: Molecular structure,
811 breakdown, genetic, bioavailability, properties and healthy and adverse effects. *Adv. Food*
812 *Nutr. Res.* **90**, 305–350 (2019).
813 40. Barba, F. J. *et al.* Bioavailability of Glucosinolates and Their Breakdown Products: Impact of
814 Processing. *Front. Nutr.* **3**, 24 (2016).
815 41. Johnson, I. T. Glucosinolates: bioavailability and importance to health. *Int. J. Vitam. Nutr.*
816 *Res. Int. Z. Vitam.- Ernahrungsforschung J. Int. Vitaminol. Nutr.* **72**, 26–31 (2002).
817 42. Sønderby, I. E., Geu-Flores, F. & Halkier, B. A. Biosynthesis of glucosinolates--gene
818 discovery and beyond. *Trends Plant Sci.* **15**, 283–290 (2010).
819 43. Pfalz, M. *et al.* Metabolic engineering in Nicotiana benthamiana reveals key enzyme
820 functions in Arabidopsis indole glucosinolate modification. *Plant Cell* **23**, 716–729 (2011).
821 44. Bureau, M. & Simon, R. JLO regulates embryo patterning and organ initiation by controlling
822 auxin transport. *Plant Signal. Behav.* **3**, 145–147 (2008).
823 45. Collin, A., Daszkowska-Golec, A. & Szarejko, I. Updates on the Role of ABSCISIC ACID
824 INSENSITIVE 5 (ABI5) and ABSCISIC ACID-RESPONSIVE ELEMENT BINDING
825 FACTORs (ABFs) in ABA Signaling in Different Developmental Stages in Plants. *Cells* **10**,

826 1996 (2021).
827 46. Moreno-Risueno, M. A. *et al.* FUSCA3 from barley unveils a common transcriptional
828 regulation of seed-specific genes between cereals and Arabidopsis. *Plant J. Cell Mol. Biol.*
829 **53**, 882–894 (2008).
830 47. Skirycz, A. *et al.* DOF transcription factor AtDof1.1 (OBP2) is part of a regulatory network
831 controlling glucosinolate biosynthesis in Arabidopsis. *Plant J. Cell Mol. Biol.* **47**, 10–24
832 (2006).
833 48. Podzimska-Sroka, D., O'Shea, C., Gregersen, P. L. & Skriver, K. NAC Transcription Factors
834 in Senescence: From Molecular Structure to Function in Crops. *Plants Basel Switz.* **4**, 412–
835 448 (2015).
836 49. Chen, Y.-F. *et al.* The WRKY6 transcription factor modulates PHOSPHATE1 expression in
837 response to low Pi stress in Arabidopsis. *Plant Cell* **21**, 3554–3566 (2009).
838 50. Li, S., Fu, Q., Huang, W. & Yu, D. Functional analysis of an Arabidopsis transcription factor
839 WRKY25 in heat stress. *Plant Cell Rep.* **28**, 683–693 (2009).
840 51. Doll, J. *et al.* Arabidopsis thaliana WRKY25 Transcription Factor Mediates Oxidative Stress
841 Tolerance and Regulates Senescence in a Redox-Dependent Manner. *Front. Plant Sci.* **10**,
842 1734 (2019).
843 52. Jiang, Y. & Deyholos, M. K. Functional characterization of Arabidopsis NaCl-inducible
844 WRKY25 and WRKY33 transcription factors in abiotic stresses. *Plant Mol. Biol.* **69**, 91–105
845 (2009).
846 53. Li, S., Fu, Q., Chen, L., Huang, W. & Yu, D. Arabidopsis thaliana WRKY25, WRKY26, and
847 WRKY33 coordinate induction of plant thermotolerance. *Planta* **233**, 1237–1252 (2011).
848 54. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from
849 Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).
850 55. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene
851 regulatory networks. *Bioinforma. Oxf. Engl.* **35**, 2159–2161 (2019).

852 56. Kishor, P. B. K., Tiozon, R. N., Fernie, A. R. & Sreenivasulu, N. Abscisic acid and its role in
853 the modulation of plant growth, development, and yield stability. *Trends Plant Sci.* **27**, 1283–
854 1295 (2022).
855 57. Bulgakov, V. P., Wu, H.-C. & Jinn, T.-L. Coordination of ABA and Chaperone Signaling in
856 Plant Stress Responses. *Trends Plant Sci.* **24**, 636–651 (2019).
857 58. Nambara, E. & Marion-Poll, A. ABA action and interactions in seeds. *Trends Plant Sci.* **8**,
858 213–217 (2003).
859 59. Lopez-Molina, L., Mongrand, S., McLachlin, D. T., Chait, B. T. & Chua, N.-H. ABI5 acts
860 downstream of ABI3 to execute an ABA-dependent growth arrest during germination. *Plant*
861 *J. Cell Mol. Biol.* **32**, 317–328 (2002).
862 60. Pollo-Rodríguez, F., Sánchez-Vicente, I. & Lorenzo, O. The turnover of ABI5 by scaffold
863 proteins to attenuate ABA signaling. *J. Exp. Bot.* **75**, 3749–3753 (2024).
864 61. Peng, J. *et al.* COP1 positively regulates ABA signaling during Arabidopsis seedling growth
865 in darkness by mediating ABA-induced ABI5 accumulation. *Plant Cell* **34**, 2286–2308
866 (2022).
867 62. Qi, L. *et al.* PHYTOCHROME-INTERACTING FACTORS Interact with the ABA Receptors
868 PYL8 and PYL9 to Orchestrate ABA Signaling in Darkness. *Mol. Plant* **13**, 414–430 (2020).
869 63. Xiong, S.-X. *et al.* The transcription factors MS188 and AMS form a complex to activate the
870 expression of CYP703A2 for sporopollenin biosynthesis in Arabidopsis thaliana. *Plant J.*
871 *Cell Mol. Biol.* **88**, 936–946 (2016).
872 64. Wang, K. *et al.* The Regulation of Sporopollenin Biosynthesis Genes for Rapid Pollen Wall
873 Formation. *Plant Physiol.* **178**, 283–294 (2018).
874 65. Zhang, Z.-B. *et al.* Transcription factor AtMYB103 is required for anther development by
875 regulating tapetum development, callose dissolution and exine formation in Arabidopsis.
876 *Plant J. Cell Mol. Biol.* **52**, 528–538 (2007).
877 66. Gerlt, J. A., Babbitt, P. C., Jacobson, M. P. & Almo, S. C. Divergent evolution in enolase

878 superfamily: strategies for assigning functions. *J. Biol. Chem.* **287**, 29–34 (2012).
879 67. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends*
880 *Plant Sci.* **19**, 698–708 (2014).
881 68. Jeffery, C. J. Current successes and remaining challenges in protein function prediction.
882 *Front. Bioinforma.* **3**, 1222182 (2023).
883 69. Jeffery, C. J. Enzymes, pseudoenzymes, and moonlighting proteins: diversity of function in
884 protein superfamilies. *FEBS J.* **287**, 4141–4149 (2020).
885 70. Jacobson, M. P., Kalyanaraman, C., Zhao, S. & Tian, B. Leveraging structure for enzyme
886 function prediction: methods, opportunities and challenges. *Trends Biochem. Sci.* **39**, 363–
887 371 (2014).
888 71. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and
889 Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2015).
890 72. Zainal-Abidin, R.-A., Harun, S., Vengatharajuloo, V., Tamizi, A.-A. & Samsulrizal, N. H.
891 Gene Co-Expression Network Tools and Databases for Crop Improvement. *Plants Basel*
892 *Switz.* **11**, 1625 (2022).
893 73. Proost, S. & Mutwil, M. CoNekT: an open-source framework for comparative genomic and
894 transcriptomic network analyses. *Nucleic Acids Res.* **46**, W133–W140 (2018).
895 74. De Bodt, S., Hollunder, J., Nelissen, H., Meulemeester, N. & Inzé, D. CORNET 2.0:
896 integrating plant coexpression, protein-protein interactions, regulatory interactions, gene
897 associations and functional annotations. *New Phytol.* **195**, 707–720 (2012).
898 75. Zheng, S. *et al.* Pupylation-Based Proximity Labeling Unravels a Comprehensive Protein
899 and Phosphoprotein Interactome of the Arabidopsis TOR Complex. *Adv. Sci. Weinh. Baden-*
900 *Wurtt. Ger.* e2414496 (2025) doi:10.1002/advs.202414496.
901 76. A., S. & R., S. A systematic review of Explainable Artificial Intelligence models and
902 applications: Recent developments and future trends. *Decis. Anal. J.* **7**, 100230 (2023).
903 77. Vivian Li, W. & Li, Y. scLink: Inferring Sparse Gene Co-expression Networks from Single-

904 cell Expression Data. *Genomics Proteomics Bioinformatics* **19**, 475–492 (2021).
905 78. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering
906 clusters in large spatial databases with noise. in *Proceedings of the Second International*
907 *Conference on Knowledge Discovery and Data Mining* 226–231 (AAAI Press, Portland,
908 Oregon, 1996).
909 79. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at
910 https://doi.org/10.48550/arXiv.1109.2378 (2011).
911 80. Brown, T. B. *et al.* Language models are few-shot learners. in *Proceedings of the 34th*
912 *International Conference on Neural Information Processing Systems* 1877–1901 (Curran
913 Associates Inc., Red Hook, NY, USA, 2020).
914 81. DeepSeek-AI *et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via
915 Reinforcement Learning. Preprint at https://doi.org/10.48550/arXiv.2501.12948 (2025).
916 82. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
917 83. Muffato, M. *et al.* Reconstruction of hundreds of reference ancestral genomes across the
918 eukaryotic kingdom. *Nat. Ecol. Evol.* **7**, 355–366 (2023).
919 84. Garcia-Hernandez, M. *et al.* TAIR: a resource for integrated Arabidopsis data. *Funct. Integr.*
920 *Genomics* **2**, 239–253 (2002).
921 85. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
922 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
923 86. Lim, P. K. *et al.* Bacteria.guru: Comparative Transcriptomics and Co-Expression Database
924 for Bacterial Pathogens. *J. Mol. Biol.* **434**, 167380 (2022).
925 87. Villanueva, E. M. F. *et al.* Protist.guru: A Comparative Transcriptomics Database for
926 Protists. *J. Mol. Biol.* **434**, 167502 (2022).
927 88. Lim, P. K., Wang, R. & Mutwil, M. LSTrAP-denovo: Automated Generation of Transcriptome
928 Atlases for Eukaryotic Species Without Genomes. *Physiol. Plant.* **176**, e14407 (2024).
929 89. Monteiro, P. T. *et al.* YEASTRACT+: a portal for cross-species comparative genomics of

930 transcription regulation in yeasts. *Nucleic Acids Res.* **48**, D642–D649 (2020).
931 90. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory
932 interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
933 91. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse
934 transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
935 92. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
936 93. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in
937 Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, D262–D266 (2004).
938 94. Wong, E. D. *et al.* Saccharomyces genome database update: server architecture, pan-
939 genome nomenclature, and external resources. *Genetics* **224**, iyac191 (2023).
940 95. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways.
941 *Brief. Bioinform.* **20**, 1085–1093 (2019).
942 96. Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete
943 human genome. *Genome Biol.* **6**, R2 (2005).
944 97. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
945 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300
946 (1995).
947 98. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set
948 enrichment analysis in Python. *Bioinforma. Oxf. Engl.* **39**, btac757 (2023).
949 99. Spearman, C. The proof and measurement of association between two things. *Int. J.*
950 *Epidemiol.* **39**, 1137–1150 (2010).
951 100. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
952 genomics. *Genome Biol.* **20**, 238 (2019).

953

Acknowledgement

The authors wish to thank SingAREN (Singapore Advanced Research and Education Network) and Mr. Soh Hwee Jin, Melvin from the High Performance Computing Centre (HPCC) of Nanyang Technological University (NTU) for providing the network infrastructure used for downloading public sequencing data. The authors also thank Asst. Prof. Jarkko Tapani Salojarvi and Asst. Prof. Goh Wen Bin Wilson from the School of Biological Sciences and Lee Kong Chian School of Medicine, respectively, of NTU, for their input in the project. The authors also thank E. Haswell (https://elizabethhaswell.carrd.co) for her help with proofreading the manuscript. Lastly, the authors thank all members of the Mutwil Lab for their support in the project and comments on the manuscript. P.K.L. would like to thank Miss Lim Rui Qi, Audrey for her help in extracting primary transcripts from mRNA isoforms and her encouragement throughout this project. P.K.L. would like to thank Asst. Prof. Anni Zhang from the School of Biological Sciences for her administrative help in his candidature. P.K.L. and S.C.L. are supported by the NTU Research Scholarship (RSS). M.M. is supported by Novo Nordisk Starting Grant.

Author information

Authors and Affiliations

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

Peng ken Lim, Ruoxi Wang, Jenet Princy Antony Velankanni and Shan Chun Lim.

Department of Plant & Environmental Sciences, University of Copenhagen, Frederiksberg C 1871, Denmark

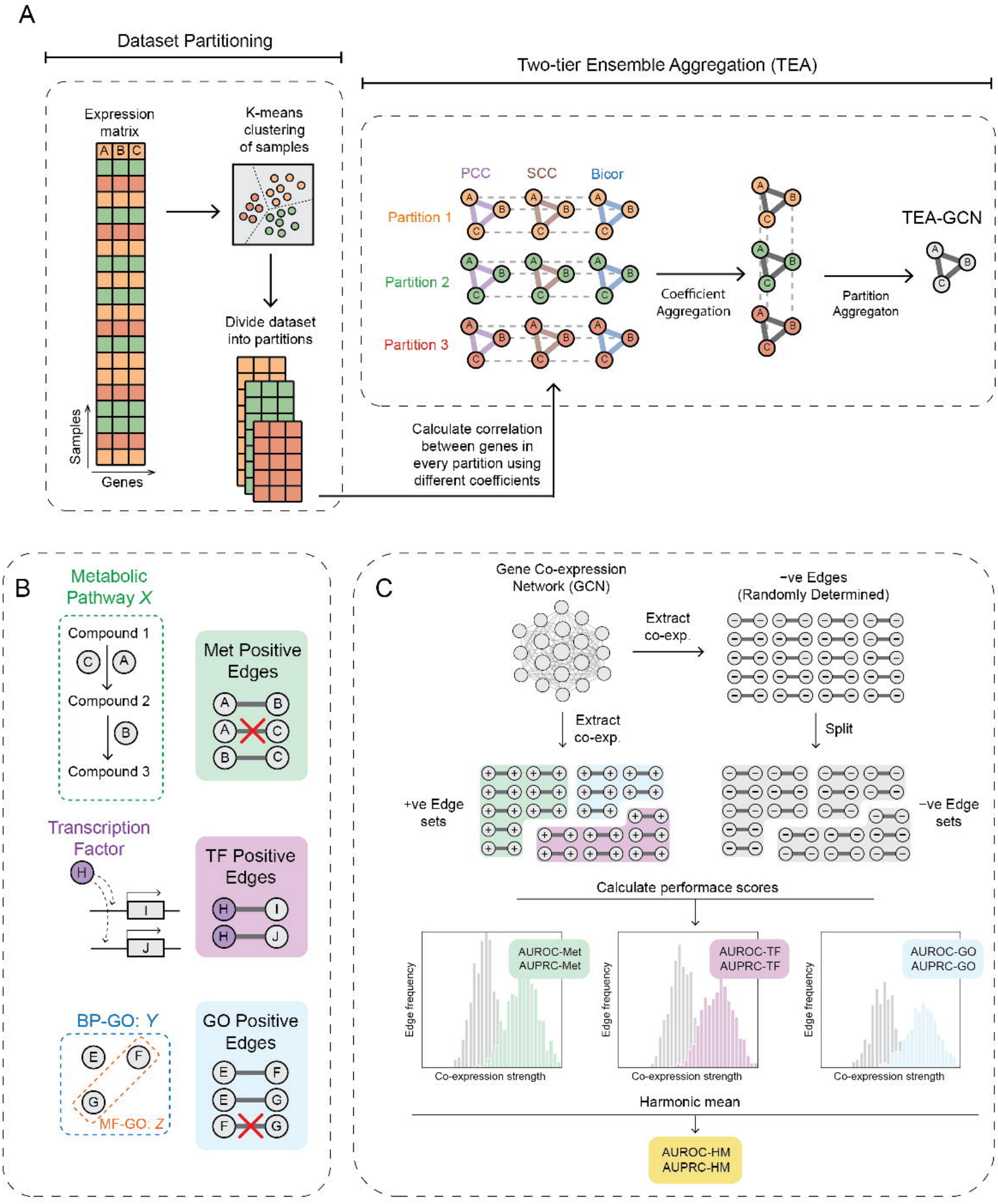
Marek Mutwil.

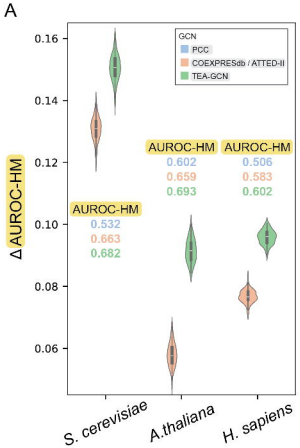
Contributions

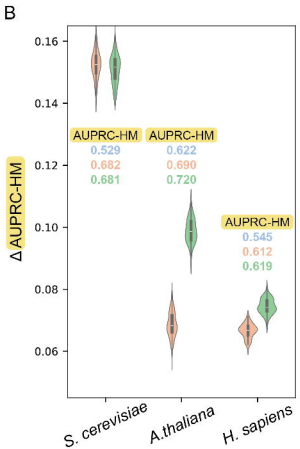
P.K.L. and M.M. conceived the project and wrote the manuscript. P.K.L. developed and wrote code for the TEA-GCN. P.K.L. and R.W. performed data analysis and generated figures. P.K.L. and S.C.L. developed, deployed the Plant-GCN database. S.C.L. is the maintainer for the Plant-GCN database. J.P.A.V. helped with the construction of edges. P.K.L. and S.C.L. were supervised by M.M., while R.W. and J.P.A. were supervised by P.K.L.. M.M. secured funding for the project.

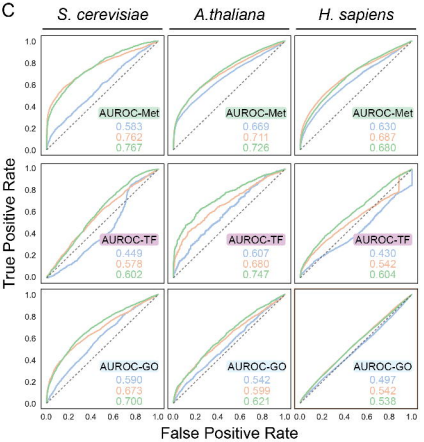
Corresponding authors

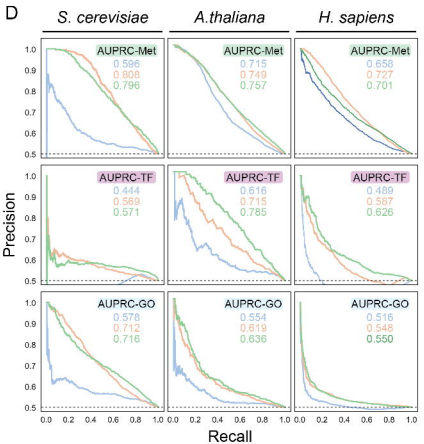
Correspondence to Marek Mutwil or Peng Ken Lim.

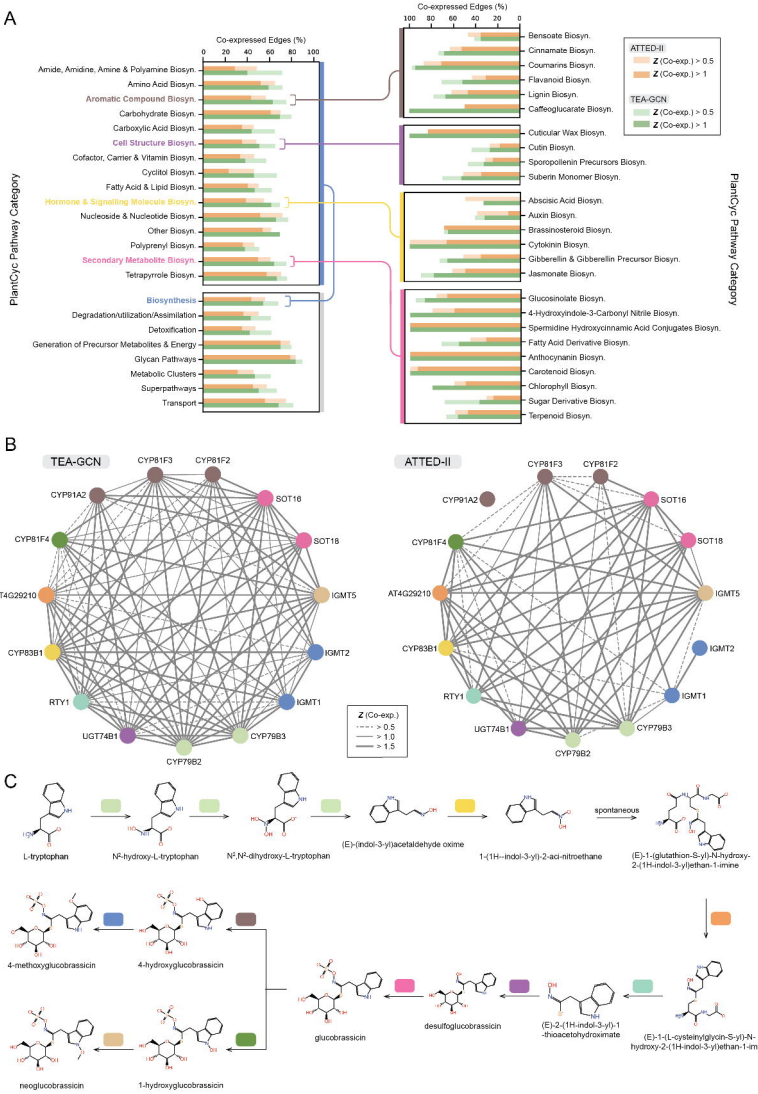

A
Dataset Partitioning
Two-tier Ensemble Aggregation (TEA)
Expression matrix
A B C
Samples
Genes
K-means clustering of samples
Divide dataset into partitions
PCC
SCC
Bicor
Partition 1
Partition 2
Partition 3
Coefficient Aggregation
Partition Aggregaton
TEA-GCN
Calculate correlation between genes in every partition using different coefficients
B
Metabolic Pathway X
Compound 1
Compound 2
Compound 3
Met Positive Edges
Transcription Factor
TF Positive Edges
BP-GO: Y
MF-GO: Z
GO Positive Edges
C
Gene Co-expression Network (GCN)
Extract co-exp.
−ve Edges (Randomly Determined)
Extract co-exp.
Split
+ve Edge sets
−ve Edge sets
Calculate performace scores
AUROC-Met
AUPRC-Met
AUROC-TF
AUPRC-TF
AUROC-GO
AUPRC-GO
Edge frequency
Co-expression strength
Harmonic mean
AUROC-HM
AUPRC-HM

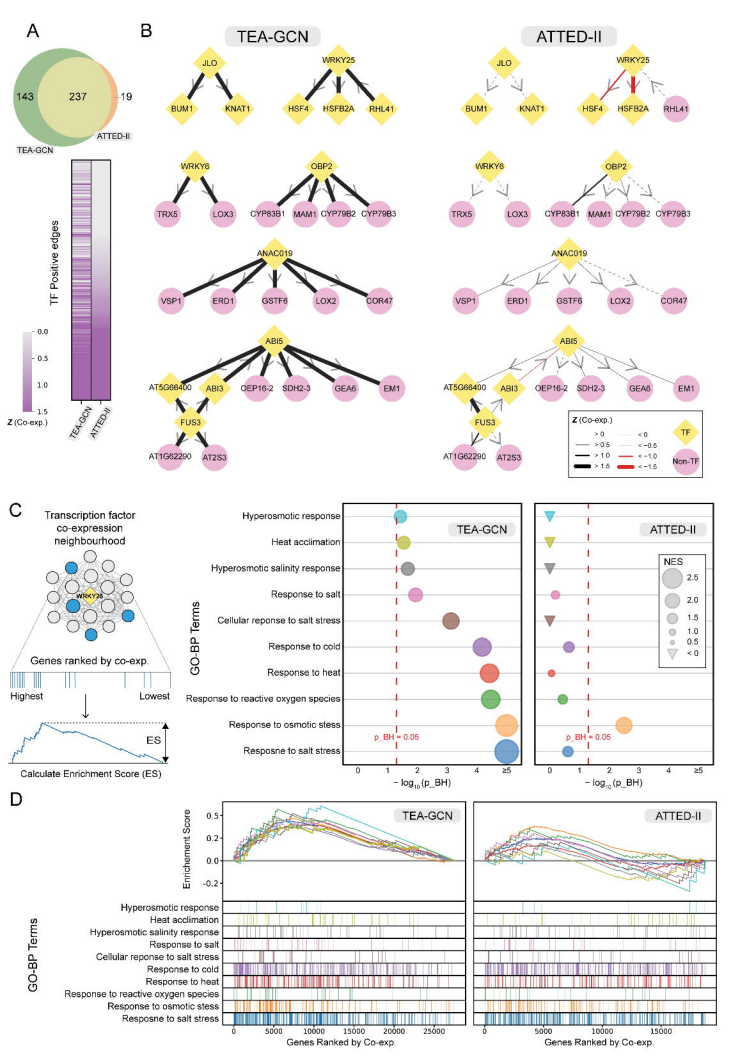

A
GCN
PCC
COEXPRESdb / ATTED-II
TEA-GCN
Δ AUROC-HM
AUROC-HM
0.532
0.663
0.682
AUROC-HM
0.602
0.659
0.693
AUROC-HM
0.506
0.583
0.602
S. cerevisiae
A.thaliana
H. sapiens

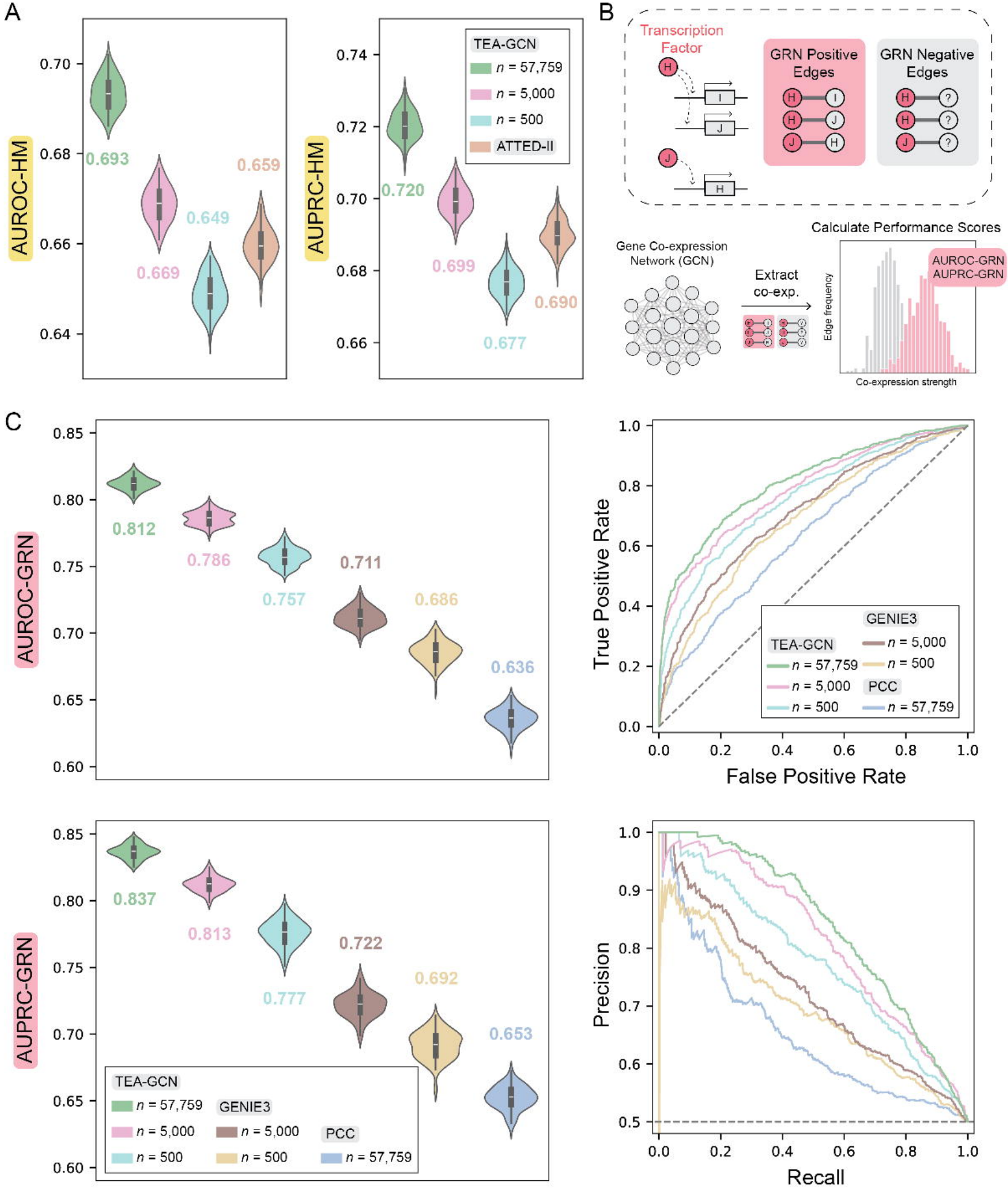

B
Δ AUPRC-HM
AUPRC-HM
0.529
0.682
0.681
AUPRC-HM
0.622
0.690
0.720
AUPRC-HM
0.545
0.612
0.619
S. cerevisiae
A.thaliana
H. sapiens

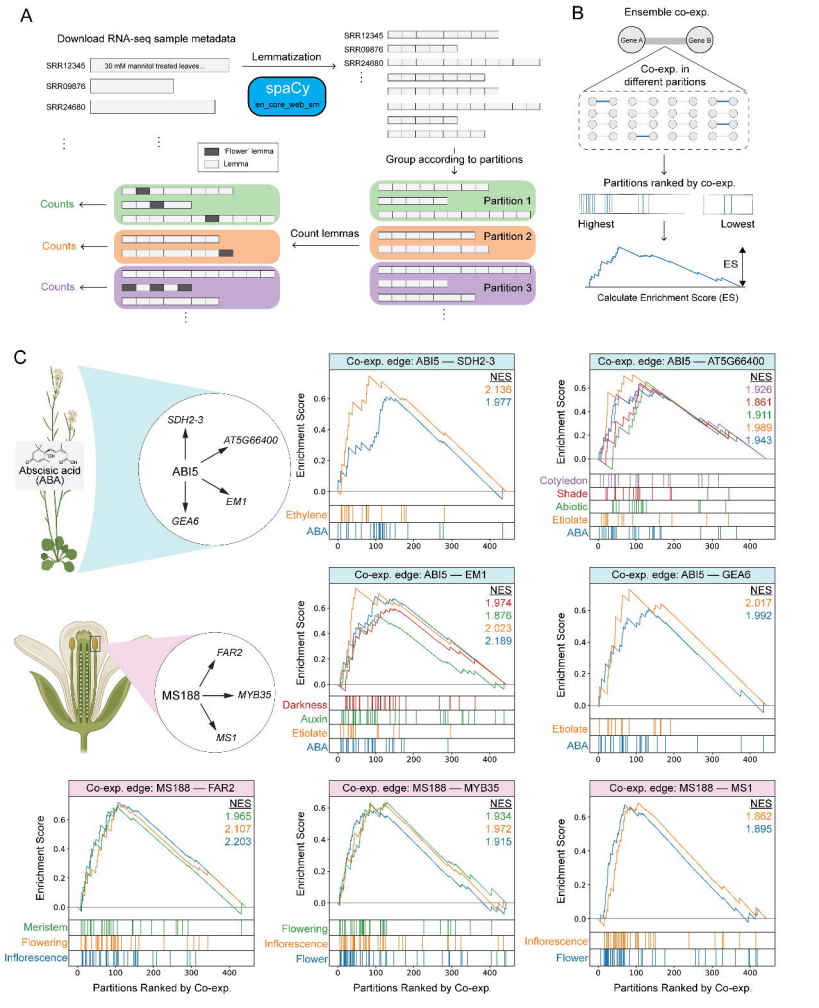

C
S. cerevisiae
A.thaliana
H. sapiens
True Positive Rate
False Positive Rate
AUROC-Met
0.583
0.762
0.767
AUROC-Met
0.669
0.721
0.726
AUROC-Met
0.630
0.687
0.680
AUROC-TF
0.449
0.578
0.602
AUROC-TF
0.607
0.680
0.747
AUROC-TF
0.430
0.542
0.604
AUROC-GO
0.590
0.673
0.700
AUROC-GO
0.542
0.599
0.621
AUROC-GO
0.497
0.542
0.538

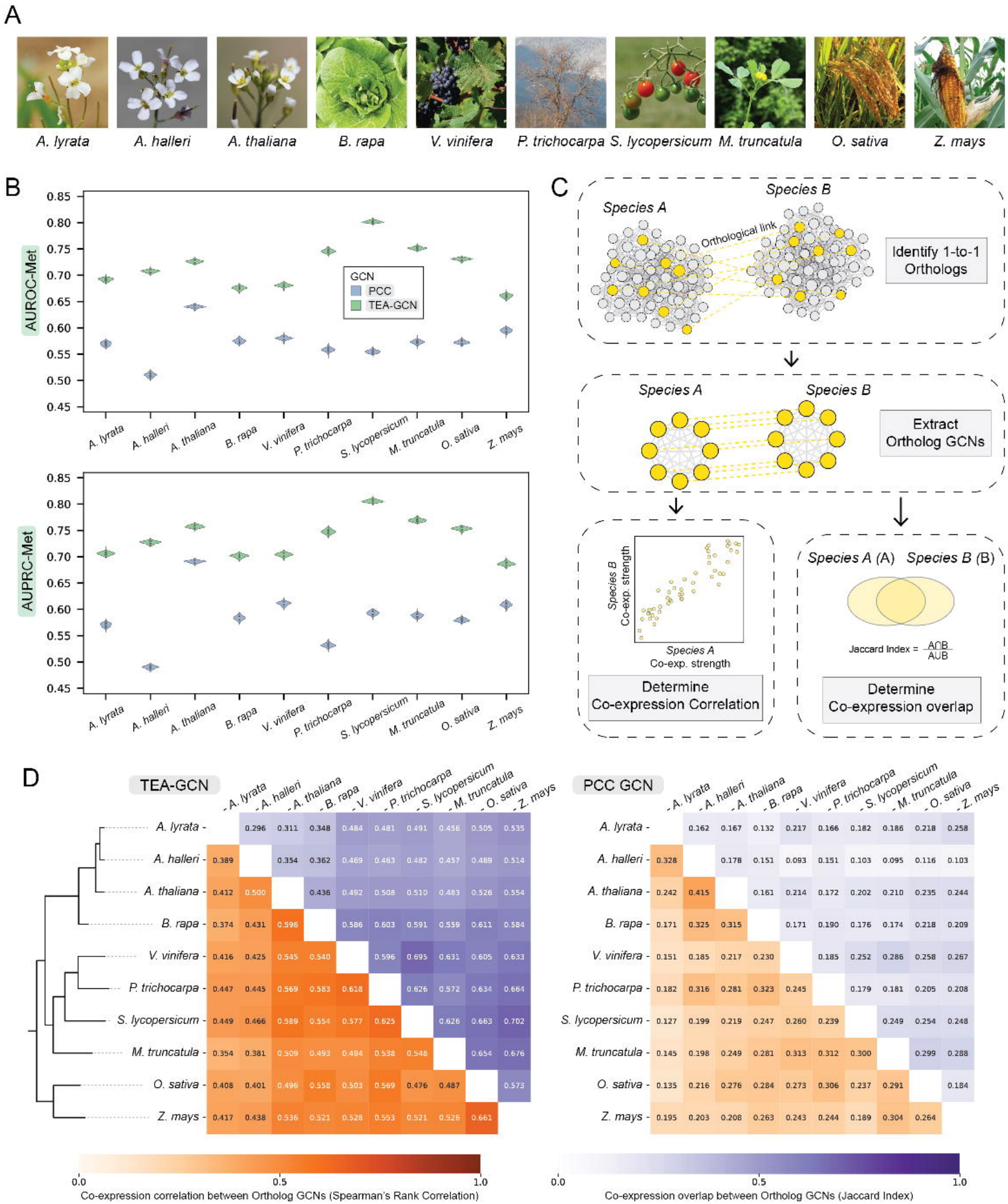

D
S. cerevisiae
A.thaliana
H. sapiens
Precision
Recall
AUPRC-Met
0.556
0.808
0.796
AUPRC-Met
0.715
0.749
0.757
AUPRC-Met
0.658
0.727
0.701
AUPRC-TF
0.444
0.569
0.571
AUPRC-TF
0.616
0.715
0.785
AUPRC-TF
0.469
0.587
0.626
AUPRC-GO
0.578
0.712
0.716
AUPRC-GO
0.554
0.619
0.636
AUPRC-GO
0.516
0.548
0.550


A
Co-expressed Edges (%)
0 20 40 60 80 100
PlantCyc Pathway Category
Amide, Amidine, Amine & Polyamine Biosyn.
Amino Acid Biosyn.
Aromatic Compound Biosyn.
Carbohydrate Biosyn.
Carboxylic Acid Biosyn.
Cell Structure Biosyn.
Cofactor, Carrier & Vitamin Biosyn.
Cyclitol Biosyn.
Fatty Acid & Lipid Biosyn.
Hormone & Signalling Molecule Biosyn.
Nucleoside & Nucleotide Biosyn.
Other Biosyn.
Polyprenyl Biosyn.
Secondary Metabolite Biosyn.
Tetrapyrrole Biosyn.
Biosynthesis
Degradation/utilization/Assimilation
Detoxification
Generation of Precursor Metabolites & Energy
Glycan Pathways
Metabolic Clusters
Superpathways
Transport
Co-expressed Edges (%)
100 80 60 40 20 0
Benzoate Biosyn.
Cinnamate Biosyn.
Coumarins Biosyn.
Flavanoid Biosyn.
Lignin Biosyn.
Caffeoglucarate Biosyn.
Cuticular Wax Biosyn.
Cutin Biosyn.
Sporopollenin Precursors Biosyn.
Suberin Monomer Biosyn.
Abscisic Acid Biosyn.
Auxin Biosyn.
Brassinosteroid Biosyn.
Cytokinin Biosyn.
Gibberellin & Gibberellin Precursor Biosyn.
Jasmonate Biosyn.
Glucosinolate Biosyn.
4-Hydroxyindole-3-Carbonyl Nitrile Biosyn.
Spermidine Hydroxycinnamic Acid Conjugates Biosyn.
Fatty Acid Derivative Biosyn.
Anthocyanin Biosyn.
Carotenoid Biosyn.
Chlorophyll Biosyn.
Sugar Derivative Biosyn.
Terpenoid Biosyn.
ATTED-II
Z (Co-exp.) > 0.5
Z (Co-exp.) > 1
TEA-GCN
Z (Co-exp.) > 0.5
Z (Co-exp.) > 1
PlantCyc Pathway Category
B
TEA-GCN
CYP81F3
CYP81F2
CYP91A2
SOT16
CYP81F4
SOT18
AT4G29210
IGMT5
CYP83B1
IGMT2
RTY1
IGMT1
UGT74B1
CYP79B3
CYP79B2
Z (Co-exp.)
> 0.5
> 1.0
> 1.5
ATTED-II
CYP81F3
CYP81F2
CYP91A2
SOT16
CYP81F4
SOT18
AT4G29210
IGMT5
CYP83B1
IGMT2
RTY1
IGMT1
UGT74B1
CYP79B3
CYP79B2
C
L-tryptophan
N⁶-hydroxy-L-tryptophan
N⁶,N⁶-dihydroxy-L-tryptophan
(E)-(indol-3-yl)acetaldehyde oxime
1-(1H--indol-3-yl)-2-aci-nitroethane
spontaneous
(E)-1-(glutathion-S-yl)-N-hydroxy-2-(1H-indol-3-yl)ethan-1-imine
(E)-1-(L-cysteinylglycin-S-yl)-N-hydroxy-2-(1H-indol-3-yl)ethan-1-imine
(E)-2-(1H-indol-3-yl)-1-thioacetohydroximate
desulfoglucobrassicin
glucobrassicin
4-methoxyglucobrassicin
4-hydroxyglucobrassicin
neoglucobrassicin
1-hydroxyglucobrassicin


A
143
237
19
TEA-GCN
ATTED-II
TF Positive edges
0.0
0.5
1.0
1.5
Z (Co-exp.)
TEA-GCN
ATTED-II
B
TEA-GCN
ATTED-II
JLO
BUM1
KNAT1
WRKY25
HSF4
HSFB2A
RHL41
WRKY6
TRX5
LOX3
OBP2
CYP83B1
MAM1
CYP79B2
CYP79B3
ANAC019
VSP1
ERD1
GSTF6
LOX2
COR47
ABI5
AT5G66400
ABI3
OEP16-2
SDH2-3
GEA6
EM1
FUS3
AT1G62290
AT2S3
Z (Co-exp.)
> 0
> 0.5
> 1.0
> 1.5
< 0
< −0.5
< −1.0
< −1.5
TF
Non-TF
C
Transcription factor co-expression neighbourhood
WRKY25
Genes ranked by co-exp.
Highest
Lowest
ES
Calculate Enrichment Score (ES)
GO-BP Terms
Hyperosmotic response
Heat acclimation
Hyperosmotic salinity response
Response to salt
Cellular reponse to salt stress
Response to cold
Response to heat
Response to reactive oxygen species
Response to osmotic stess
Response to salt stress
TEA-GCN
ATTED-II
NES
2.5
2.0
1.5
1.0
0.5
< 0
p_BH = 0.05
0
1
2
3
4
≥5
− log10 (p_BH)
D
Enrichment Score
0.5
0.2
0.0
-0.2
TEA-GCN
ATTED-II
GO-BP Terms
Hyperosmotic response
Heat acclimation
Hyperosmotic salinity response
Response to salt
Cellular reponse to salt stress
Response to cold
Response to heat
Response to reactive oxygen species
Response to osmotic stess
Response to salt stress
0
5000
10000
15000
20000
25000
Genes Ranked by Co-exp.


A
AUROC-HM
0.70
0.68
0.66
0.64
0.693
0.669
0.649
0.659
AUPRC-HM
0.74
0.72
0.70
0.68
0.66
0.720
0.699
0.677
0.690
TEA-GCN
n = 57,759
n = 5,000
n = 500
ATTED-II
B
Transcription Factor
GRN Positive Edges
GRN Negative Edges
Gene Co-expression Network (GCN)
Extract co-exp.
Calculate Performance Scores
AUROC-GRN
AUPRC-GRN
Edge frequency
Co-expression strength
C
AUROC-GRN
0.85
0.80
0.75
0.70
0.65
0.60
0.812
0.786
0.757
0.711
0.686
0.636
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
GENIE3
TEA-GCN
n = 5,000
n = 57,759
n = 500
n = 5,000
PCC
n = 500
n = 57,759
AUPRC-GRN
0.837
0.813
0.777
0.722
0.692
0.653
TEA-GCN
n = 57,759
n = 5,000
n = 500
GENIE3
n = 5,000
n = 500
PCC
n = 57,759
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5


A
Download RNA-seq sample metadata
SRR12345
30 mM mannitol treated leaves...
SRR09876
SRR24680
Lemmatization
spaCy
en_core_web_sm
SRR12345
SRR09876
SRR24680
'Flower' lemma
Lemma
Group according to partitions
Counts
Counts
Counts
Count lemmas
Partition 1
Partition 2
Partition 3
B
Ensemble co-exp.
Gene A
Gene B
Co-exp. in different paritions
Partitions ranked by co-exp.
Highest
Lowest
ES
Calculate Enrichment Score (ES)
C
Abscisic acid (ABA)
SDH2-3
AT5G66400
ABI5
EM1
GEA6
MS188
FAR2
MYB35
MS1
Co-exp. edge: ABI5 — SDH2-3
NES
2.136
1.977
Ethylene
ABA
Co-exp. edge: ABI5 — AT5G66400
NES
1.926
1.861
1.911
1.989
1.943
Cotyledon
Shade
Abiotic
Etiolate
ABA
Co-exp. edge: ABI5 — EM1
NES
1.974
1.876
2.023
2.189
Darkness
Auxin
Etiolate
ABA
Co-exp. edge: ABI5 — GEA6
NES
2.017
1.992
Etiolate
ABA
Co-exp. edge: MS188 — FAR2
NES
1.965
2.107
2.203
Meristem
Flowering
Inflorescence
Co-exp. edge: MS188 — MYB35
NES
1.934
1.972
1.915
Flowering
Inflorescence
Flower
Co-exp. edge: MS188 — MS1
NES
1.862
1.895
Inflorescence
Flower
Enrichment Score
Partitions Ranked by Co-exp.


A
A. lyrata
A. halleri
A. thaliana
B. rapa
V. vinifera
P. trichocarpa
S. lycopersicum
M. truncatula
O. sativa
Z. mays
B
AUROC-Met
AUPRC-Met
GCN
PCC
TEA-GCN
C
Species A
Species B
Orthological link
Identify 1-to-1 Orthologs
Extract Ortholog GCNs
Species B Co-exp. strength
Species A Co-exp. strength
Determine Co-expression Correlation
Species A (A)
Species B (B)
Jaccard Index = A∩B / A∪B
Determine Co-expression overlap
D
TEA-GCN
PCC GCN
Co-expression correlation between Ortholog GCNs (Spearman's Rank Correlation)
Co-expression overlap between Ortholog GCNs (Jaccard Index)