TITLE: Constructing Gene Co-functional and Co-regulatory Networks from Public Transcriptomes using 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) can reveal useful gene co-functional and co-regulatory relationships. However, current GCN construction methodologies are sensitive to batch effects and sample composition, limiting their performance in generating GCNs from public RNA-seq samples abundant for many species. Here, we detail the development of TEA-GCN (Two-Tier Ensemble Aggregation-GCN), a GCN construction method that leverages unsupervised transcriptomic dataset partitioning and multi-metric co-expression scoring to derive ensemble gene co-expression. Benchmarking over 450,000 public RNA-seq samples across 12 species, TEA-GCN outperforms the state-of-the-art in predicting gene functions and inferring gene regulatory networks. Through the use of natural language processing, we also show that the biologically-relevant dataset partitions with high co-expression can identify tissue-/condition-specific co-expression in TEA-GCN, providing an unprecedented level of explainability. Furthermore, we show that TEA-GCNs exhibit enhanced conservation across species, making them suitable for multi-species comparative studies. TEA-GCN is available at https://github.com/pengkenlim/TEA-GCN.

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[5,6]. The rapidly declining costs of RNA-seq, coupled with the fast rate at which new genomes are published, have 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 Subagging 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 this work, we demonstrate and characterize the predictive performance of GCNs generated using the Two-Tier Ensemble Aggregation GCN (TEA-GCN) method, which 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 the 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 consensus on which is the most ideal for generating robust GCNs, as different coefficients excel at 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 in dataset partitions based on all three coefficients using the Maximum (Max) function (see Supplemental methods), which effectively reports correlation strengths according to the coefficient that best describes co-expression relationships.

Partition Aggregation aims to capture tissue/condition-specific co-expression between functionally related genes that might be missed if all samples are used for the network construction. For example, when we consider all 71,720 *A. thaliana* RNA-seq samples together, the functionally-related enzyme pair *AT4G37400–AT2G20610* appears essentially uncorrelated (PCC = 0.016, SCC = –0.066, bicor = 0.094; Figure S1). After k-means partitioning of the same samples, however, individual partitions reveal strong condition-specific co-expression for this pair (e.g., SCC up to ~0.59 and bicor up to ~0.82 in some partitions, while others remain weakly correlated). TEA-GCN exploits precisely this pattern by learning from partition-specific correlations and then aggregating them, enabling it to recover regulatory relationships that are

119 invisible to conventional GCNs computed on the global compendium. Consistent with this aim, in
120 *A. thaliana,* we demonstrate that the k-means partitions show strong agreement with independent
121 and manually-curated EVOREPRO tissue labels[27] via the use of V-measure[28] , which is an
122 entropy-based clustering metric to measure the similarity between clustering labels and ground
123 truth labels. We found the tissue labels to be significantly and repeatedly more similar to the k-
124 means partitions (V-measure of 0.697, Figure S2) than randomly permuted partitions (V-measure
125 of 0.508 ± 0.012 [median ± IQR]), indicating that the partitions indeed correspond to distinct
126 organ/condition contexts.

127 The TEA-GCN framework uses K-means clustering[29], an unsupervised clustering
128 algorithm, to partition the transcriptomic dataset into samples with similar transcriptome
129 profiles[30,31] (Figure 1a). In addition, we systematically evaluated hierarchical agglomerative
130 clustering[32] and DBSCAN[33] as alternative partitioning strategies (Figure S3). Hierarchical
131 clustering produced very similar, but on average slightly lower, AUROC-HM and AUPRC-HM
132 values compared with K-means (Figure S3), whereas DBSCAN achieved slightly higher peak
133 performance for specific combinations of epsilon and minimum samples, but was strongly
134 dependent on these hyperparameters (Figure S3a), resulting in unstable numbers of clusters,
135 silhouette coefficients[34] and noise ratios across settings (Figure S4). Because K-means yielded
136 consistently high TEA-GCN performance together with stable cluster sizes and noise levels
137 across clustering granularities and species, we adopted K-means as the default clustering method
138 in TEA-GCN. Through Partition Aggregation, the co-expression strengths between all gene pairs
139 are aggregated across all partitions to detect tissue/condition-specific co-expression using the
140 Rectified Average (RAvg) function, where the negative co-expression strengths from every
141 dataset partition are converted to zero before averaging, resulting in more accurate co-expression
142 determination (supplemental results, Figure S5). The resultant ensemble co-expression from two
143 rounds of aggregation was then subjected to Mutual Rank Transformation (supplementary
144 methods) to yield the final TEA-GCN.

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

146 We benchmarked TEA-GCN using 150,096, 71,720, and 95,553 public RNA-seq samples for *S.*
147 *cerevisiae*, *A. thaliana*, and *H. sapiens,* respectively (see Methods), and evaluated the ability of
148 networks to recover experimentally-verified metabolic networks (Met) (Figure 1b; green),
149 transcription factor-target relationships (TF) (Figure 1b; purple), and genes involved in the same
150 biological process (GO) (Figure 1b; blue).

151 We found that the combination of Coefficient Aggregation and Partition Aggregation in
152 TEA-GCN results in a dramatic increase in GCN performance for *S. cerevisiae*, *A. thaliana*, and
153 *H. sapiens* (Supplemental Results, Figure S5). To compare its performance to the state-of-the-
154 art approaches, we compare the performances of *S. cerevisiae*, *A. thaliana*, and *H. sapiens* GCNs
155 from two industry-leading databases that use the Subagging approach, COXPRESdb[15]
156 (https://coxpresdb.jp/download/) and ATTED-II[18] (https://atted.jp/download/). As expected, these
157 Subagging ensemble GCNs have a better overall performance than non-ensemble PCC GCNs
158 across all three species (ΔAUROC-HM and ΔAUPRC-HM > 0.06-0.16) (Figure 2a and 2b;
159 Supplementary Data 1). However, TEA-GCNs showed significantly (p-value < 0.05) higher
160 performance for nearly all metrics (Figures 2a, 2b, and S6; Supplementary Data 1) as compared
161 to Subagging ensemble GCNs. With square brackets denoting 95% confidence interval range,
162 these improvements correspond to as much as 0.0304 [0.0293, 0.321] and 0.0340 [0.325, 0.355]
163 in AUROC-HM and AUROC-PRC, respectively (Figure 2a and b).

164 To provide a more fine-grained comparison, we measured the performance differences in
165 predicting gene regulatory relationships (TF), co-involvement in metabolic pathways (Met), and
166 biological processes (GO-BP). TEA-GCNs were consistently more performant in retrieving the
167 experimentally verified gene regulatory relationships (TF) across the three species (Figure 2c and
168 2d; Supplementary Data 1). For example, *A. thaliana* TEA-GCN has a TF performance (AUROC-
169 TF, AUPRC-TF) of (0.749 0.786) as compared to (0.680, 0.715) for *A. thaliana* ATTED-II GCN,
170 representing an AUROC-TF and AUPRC-TF improvement of 0.0691 [0.0660, 0.0715] and 0.0709
171 [0.0677,0.0773], respectively; *H. sapiens* TEA-GCN has a TF performance of (0.604, 0.626) as
172 compared to (0.542, 0.587) for *H. sapiens* COXPRESdb GCN, representing an AUROC-TF and
173 AUPRC-TF improvement of 0.0625 [0.0613, 0.0639] and 0.0388 [0.0368, 0.0407], respectively
174 (Figure 2c and 2d; Supplementary Data 1). Conversely, the ATTED-II/COXPRESdb GCNs are
175 marginally better at identifying pathway co-involvement of enzymes for *S. cerevisiae* and *H.*
176 *sapiens*, while *A. thaliana* TEA-GCN has a slightly higher Met performance than *A. thaliana*
177 ATTED-II, representing an AUROC-Met and AUPRC-Met improvement of 0.0125 [0.0117,
178 0.0135] and 0.0069 [0.0059, 0.0082], respectively. TEA-GCN is also better than
179 COXPRESdb/ATTED-II GCNs at identifying genes sharing the same GO-BP term in *S. cerevisiae*
180 (AUROC-GO improvement: 0.0264 [0.0246, 0.0278], AUPRC-GO improvement: 0.0035 [0.0015,
181 0.0058]) and in *A.thaliana* (AUROC-GO improvement: 0.0227 [0.0208, 0.0253], AUPRC-GO
182 improvement: 0.0183 [0.0165, 0.0206]) (Figure 2c and 2d; Supplementary Data 1). We speculate
183 that the superior performance of TEA-GCN lies in its ability to detect condition/tissue-specific co-

184 expression from individual dataset partitions. The observed TF performance disparity between
185 TEA-GCN and the Subagging method is more pronounced in multicellular *A. thaliana* and *H.*
186 *sapiens* than in unicellular *S. cerevisiae* (Figure 2c and 2d; Supplementary Data 1), which could
187 be attributed to the presence of more specialized transcriptional programs in multicellular
188 organisms than unicellular organisms[35,36]. Since each measurement of performance score is
189 calculated from a set of randomly generated negative edges, we found that the variances of
190 performance scores are largely influenced by underlying positive and negative edges used for
191 different aspects and organisms, rather than the evaluated networks. For example, the
192 AUPRC/AUROC-TF scores of *A. thaliana* GCNs exhibit a relatively higher (≥ 0.01) interquartile
193 range (IQR) (Supplementary Data 1) as compared to the lower IQRs (≤ 0.004) of
194 AUPRC/AUROC-Met scores from *H. sapiens* GCNs (Supplementary Data 1).

195 TEA-GCN captures co-expression between experimentally validated enzymes in 196 *Arabidopsis thaliana* metabolic pathways

197 To characterize how well the TEA-GCN framework captures co-expression between enzymes of
198 the same pathway, we compared the co-expression strengths of Met positive edges in *A. thaliana*
199 TEA-GCN and ATTED-II GCN. We define Met positive edges as edges that connect
200 experimentally validated enzymes within individual PlantCyc pathways[37]. For a fair comparison of
201 co-expression strengths across the two CGNs, we used standardized co-expression strengths
202 (Z(Co-exp.)), which measure the co-expression strengths of each edge relative to the co-
203 expression strengths of all edges in the GCN (see Methods).

204 *A. thaliana* TEA-GCN performed better in identifying Met positive edges than the ATTED-
205 II GCN across all major pathway classes (Figure 3a; Grey section). The only exception was the
206 "Generation of Precursor Metabolites & Energy" class, where performance between the two
207 GCNs is nearly identical (Figure 3a; Supplementary Data 2), irrespective of co-expression
208 thresholds (i.e., Z(Co-exp.) ≥ 0.5 and Z(Co-exp.) ≥ 1). TEA-GCN excelled in the "Metabolic
209 Clusters" and "Biosynthesis" classes, where the percentage of positive edges at both thresholds
210 is more than 10% higher than in ATTED-II GCN (Figure 3a; Supplementary Data 2). The
211 "Generation of Precursor Metabolites & Energy" class consists of conserved pathways such as
212 the TCA cycle, glycolysis, respiration, and photosynthesis that are ubiquitously expressed and
213 not likely to be condition/tissue-specific, which could explain the similarity of performances in both
214 GCNs. Conversely, "Biosynthesis" and "Detoxification" classes, which tend to comprise
215 specialized pathways that are lineage-specific and/or condition/tissue-specific[38–43], have more co-

216 expressed edges in TEA-GCN than ATTED-II GCN at both thresholds. These observations are
217 compatible with our hypothesis that the good performance of TEA-GCN is due to its ability to
218 capture condition/tissue-specific co-expression.

219 To better compare the *A. thaliana* GCNs of TEA-GCN and ATTED-II, we investigated the
220 different subclasses within the "Biosynthesis" class. TEA-GCN identified at least 10% more co-
221 expressed edges than ATTED-II for "Hormone and Signaling Molecule Biosynthesis", "Cyclitol
222 Biosynthesis", "Aromatic Compound Biosynthesis", "Cell Structure Biosynthesis", "Secondary
223 Metabolite Biosynthesis", "Amide, Amidine, Amine, and Polyamine Biosynthesis", at both co-
224 expression thresholds (Figure 3a, Blue). We observed an extra 23% of positive met edges
225 belonging to "Hormone and Signaling Molecule Biosynthesis" for TEA-GCN (63.00%; 126 out of
226 200) compared to ATTED-II GCN (38.5%; 77 out of 200) (Figure 3a, blue; Supplementary Data
227 2) at the stricter co-expression threshold of Z(Co-exp.) ≥ 1. This observation is largely attributed
228 to the "Gibberellin and Gibberellin Precursor Biosynthesis" and "Jasmonate Biosynthesis" (Figure
229 3a, yellow) that collectively account for more than half of the edges belonging to "Hormone and
230 Signaling Molecule Biosynthesis" (Supplementary Data 2). Given that the biosynthesis of
231 hormones is known to be highly localized to certain tissues at specific stages of development or
232 environmental conditions[44–46], we attribute TEA-GCN's higher performance to its ability to capture
233 condition-specific co-expression. Other notable classes where TEA-GCN excels include "Fatty
234 Acid Derivative Biosynthesis" (TEA-GCN: 66.04% and ATTED-II: 31.45%, strict threshold; 159
235 total edges), "Flavonoid Biosynthesis" (TEA-GCN: 66.93% and ATTED-II: 30.71%, strict
236 threshold; 127 total edges), "Lignin Biosynthesis" (TEA-GCN: 77.24% and ATTED-II: 47.15%,
237 strict threshold; 123 total edges) and "Glucosinolate Biosynthesis" (TEA-GCN: 93.92% and
238 ATTED-II: 65.19%, strict threshold; 181 total edges) (Figure 3a; Supplementary Data 2).

239 To exemplify the utility of TEA-GCN in predicting pathway memberships of enzymes, we
240 compared the co-expression strengths between genes that code for experimentally validated
241 enzymes participating in the "Glucosinolate Biosynthesis from Tryptophan" PlantCyc pathway
242 (henceforth referred to by its corresponding PlantCyc pathway accession, PWY-601; Figure 3b
243 and 3C). There is significant research interest in transgenically modifying the level of specific
244 glucosinolates, which can be beneficial or toxic when consumed[47–49], in Brassicaceae food crops[50]
245 which relies on elucidating their biosynthetic enzymes.

246 Experimentally validated enzymes of PWY-601 tended to be co-expressed with more
247 experimentally validated enzymes from the same pathway in *A. thaliana* TEA-GCN than in

ATTED-II GCN (Figure 3b). Furthermore, genes coding for enzymes involved in the derivatization of 3-IMG to 4-MoIMG (i.e., *IGMT1*, *IGMT2*, *CYP91A2*, *CYP81F3*, and *CYP81F2*) have notably lower co-expression strengths with other PWY-601 enzymes in ATTED-II GCN, as opposed to the more connected PWY-601 network of TEA-GCN (Figure 3b and 3c). This lack of co-expression with other enzymes is especially prominent for *CYP91A2* and *IGMT2*, which codes for the monooxygenase and methyltransferase that catalyze the hydroxylation of 3-IMG to 4-hydroxylglucobrassicin (4-OhIMG) and the methylation of 4-OhIMG to 4-MoIMG, respectively[51]; *CYP91A2* is not co-expressed with any genes in PWY-601, while *IGMT2* is only connected to *IGMT1* which codes for a methyltransferase catalyzing the same reaction step (Figure 3b). All in all, we have shown that TEA-GCN can better capture co-expression between genes coding for enzymes within specialized metabolic pathways than the state-of-the-art Subagging method, possibly due to TEA-GCN's ability to detect tissue/conditional-specific gene co-expression.

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

Our analysis revealed that TEA-GCN is particularly performant in revealing transcription factors and their gene targets (Figure 2c and 2d; TF aspect). Across the curated TF benchmark, roughly one third of TF positive edges are co-expressed (Z(Co-exp.) ≥ 1) recovered by both GCNs (244/673, 36.3%), another third are recovered only by TEA-GCN (202/673, 30.0%), and only a small minority are recovered only by ATTED-II (11/673, 1.6%; Figure 4a). Thus, TEA-GCN contains nearly all TF–target edges found in ATTED-II and identifies many additional positive edges. Since TEA-GCN identifies not only more TF positive edges than ATTED-II but nearly all TF positive edges identified by ATTED-II, TEA-GCN serves well as a replacement to the Subagging method in uncovering transactivation regulatory links.

To further characterize the difference between the gene regulatory networks identified by TEA-GCN and ATTED-II, we analyzed transcription factors where TEA-GCN, but not ATTED-II, showed high co-expression (Z(Co-exp.) >1.5) with their transactivation gene targets (Figure 4b). The transcription factor genes identified in this manner are known to be highly tissue/condition-specific in their expression, which further supports our hypothesis that TEA-GCN excels in detecting tissue/condition-specific co-expression. To exemplify, *JAGGED LATERAL ORGANS*'s (*JLO*) expression is specific to embryos[52], while *ABA INSENSITIVE 5*'s (*ABI5*) and *FUSCA3*'s (*FUS3*) expression are seed-specific[53,54]. *OBF-BINDING PROTEIN 2*'s (*OBP2*) expression is induced by wounding and phytohormone Methyl jasmonate (MeJA)[55], *NAC DOMAIN CONTAINING PROTEIN 19*'s (*ANAC019*) expression is induced by drought, high-salt stress, and

280 phytohormone abscisic acid[56], and *WRKY6*'s expression is induced by phosphate-starvation[57].
281 Curiously, *WRKY25* is found by ATTED-II GCN to be negatively co-expressed with its
282 transactivation targets of *HEAT SHOCK FACTOR 4* (*HSF4*) and *HEAT SHOCK*
283 *TRANSCRIPTION FACTOR B2A* (*HSFB2A)*[58](Figure 4b). Conversely, these genes are
284 positively co-expressed in TEA-GCN (Figure 4b), which is in agreement with northern blot studies
285 showing that HSFB2A and HSF4 are more highly expressed in WRKY25 overexpressors[58].
286 *WRKY25* is involved in responses against a multitude of abiotic stresses such as heat, cold,
287 osmotic, and high-salt conditions[58–61]. Given the multimodal functions of *WRKY25* and its
288 participation in various abiotic responses, we speculate that its mode of action might be condition-
289 specific, which might be the reason why its co-expression with *HSF4* and *HSFB2A* was not
290 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[62]
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 4d, Figure S7). 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 4c and 4d). 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 4d). 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 co-expression analyses.

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

313 Given that there may be little publicly available RNA-seq data for some species, we investigated
314 TEA-GCN performance on smaller datasets. To this end, we generated GCNs for downsampled
315 datasets consisting of 5,000 samples and 500 samples that were randomly selected (without
316 replacement) from the *A. thaliana* public transcriptome dataset. While the overall performance of
317 the *A. thaliana* TEA-GCNs markedly decreases as the datasets become smaller, the *A. thaliana*
318 TEA-GCN at n = 500 is still substantially more performant (AUROC-HM: 0.649, AUPRC-HM:
319 0.677; Figure 5a; Supplementary Data 3) than the non-ensemble PCC GCN (AUROC-HM: 0.602,
320 AUPRC-HM: 0.622; Figure 2a and 2b; Supplementary Data 1) generated from the complete *A.*
321 *thaliana* public transcriptome dataset (n = 71,720). Crucially, *A. thaliana* TEA-GCN at n = 5,000
322 can still outperform ATTED-II GCN (AUROC-HM improvement: 0.0105 [0.0094, 0.0122], AUPRC-
323 HM improvement: 0.0128 [0.0116, 0.0148]; Figure 5a; Supplementary Data 1), despite using a
324 downsampled dataset that is more than one order of magnitude smaller than the complete
325 dataset. This outperformance is made more impressive considering that ATTED-II GCN is
326 generated via the Subagging method using 14,741 curated, sample-balanced, and batch-
327 corrected RNA-seq samples (according to https://atted.jp/download/), as opposed to *A. thaliana*
328 TEA-GCN at n = 5,000, which uses 5,000 non-batch-corrected RNA-seq samples that have been
329 randomly selected from the public transcriptomic dataset. Thus, the performance of TEA-GCN is
330 not only translatable to smaller datasets but can likely outperform the Subagging methods, which
331 require time-consuming and complex annotation and balancing of samples.
332
333 To further assess the impact of curation and batch correction, we additionally constructed
334 TEA-GCNs from the 14,741 *A. thaliana* curated samples used by ATTED-II (14,741', dashed blue
335 violin plot) and also applied batch correction with ComBat[63] (ComBat(14,741*), solid blue violin
336 plot). Firstly, we observed that TEA-GCN still outperforms ATTED-II GCN when built from the
337 same 14,741 RNA-seq samples (Figure 5a; AUROC-HM improvement: 0.0105 [0.0094, 0.0122]
338 AUPRC-improvement: 0.0128 [0.0116, 0.0148]). Furthermore, we observed a marginally
339 detrimental effect on TEA-GCN performance (Figure 5a; AUROC-HM decrease: 0.0038 [0.26E-
340 05, 0.0048], no significant change in AUPRC-HM) when the 14,741 RNA-seq samples were batch
341 corrected (see Methods), indicating that our method's robustness to batch effect does not require
342 prior batch correction of expression data.

TEA-GCN outperforms dedicated GRN inference methods

As demonstrated earlier, the TEA-GCN framework excels in revealing gene regulatory networks (Figures 2a, 2d, and 4), which prompted us to compare its GRN-inference capabilities to well-established GRN inference methods GENIE3[64] and GRNBoost2[65]. Comparison of TEA-GCN and GENIE3 revealed that TEA-GCN outperforms GENIE3 in identifying directional transregulatory relationships (GRN positive edges, Figure 5b, Supplementary Data 4) even when a smaller dataset comprising < 10% of samples is used (Figure 5c; Supplementary Data 5). TEA-GCN at *n* = 500 outperforms the GENIE3 GRN at *n* = 5,000 (AUROC-GRN improvement: 0.0504 [0.0478, 0.0528], AUPRC-GRN improvement: 0.0568 [0.0544, 0.0602]) and the GRNBoost2 GRN at n = 5,000 (AUROC-GRN improvement: 0.0598 [0.0578, 0.0624], AUPRC-GRN improvement: 0.0569 [0.0542, 0.0596]; Figure 5c) in GRN inference despite a ten-fold difference in dataset size. *A. thaliana* TEA-GCN generated using the complete dataset (*n* = 71,720) has the best GRN-inference ability (AUROC-GRN: 0.817, AUPRC-GRN: 0.837; Figure 5c; Supplementary Data 5). Thus, TEA-GCNs can outperform dedicated GRN algorithms, such as GENIE3 and GRNBoost2, 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[64] and GRNBoost2[65] 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, Supplementary Data 6), and heuristically annotates ensemble co-expression edges with co-expression contexts based on their underlying co-expression profile across partitions (Figure 6b; Supplementary Data 6). To test our approach, we applied this method to GRN-positive edges that exhibited strong co-expression in the *A. thaliana* TEA-GCN (i.e., Z(Co-exp.) ≥ 1.5). The majority of identified edges (326 out of 446;

375 73.09%; Supplementary Data 7) could be annotated with at least one lemma, and these
376 annotations largely correspond with the known functions of the gene pairs (Figure 6c;
377 Supplementary Data 7). To exemplify, we selected two well-characterized transcription factor
378 genes, *ABSCISIC ACID INSENSITIVE 5* (*ABI5*) and *MALE STERILE 188* (*MS188*). Five positive
379 edges of *ABI5* (connected to *GEA6*, *EM1*, *SDH2-3,* and RAB18*)* were annotated with abscisic
380 acid (ABA), a phytohormone that mediates important physiological processes such as seed
381 dormancy maintenance and abiotic stress response[66–68] via "Aba"/"Abscisic" experimental context
382 lemmas (Figure 6c; Supplementary Data 7). This is consistent with the prominent role of ABI5 in
383 regulating ABA signaling[53,69,70]. Interestingly, these *ABI5* edges were annotated with experimental
384 contexts reflecting low-light/dark growth conditions (i.e., "Shade", "Etiolate", "Darkness", "Dark";
385 Figure 6c; Supplementary Data 7), which represents novel information that can be used to further
386 dissect the regulatory wiring underpinning ABA signaling of ABI5. The revealed darkness-related
387 experimental contexts are supported by the recent identification of COP1 E3 ligase[71] and
388 phytochrome-interacting factors[72] as darkness-specific modulators of ABI5 activity.

389 In another example, all three edges of *MS188* (connected to *FAR2*, *MYB35*, and *MS1*)
390 were annotated with experimental contexts related to floral tissues (i.e. "Flower", "Floral",
391 "Inflorescence"), which is congruent with MS188's known function in anther developmental
392 processes (e.g., tapetum development, pollen exine formation, and pollen callose development;
393 Figure 6c; Supplementary Data 7)[73–75]. Thus, TEA-GCN, when combined with NLP approaches,
394 produces interpretable edges that allow the elucidation of the experimental context, further
395 boosting gene function prediction approaches.

396 TEA-GCNs Networks Are Highly Conserved Across Species

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

406 *Arabidopsis lyrata* (*A. lyrata*) and *Arabidopsis halleri (A. halleri)*, which belong to the same genus
407 as *A. thaliana*.

408 To benchmark GCN performance, we defined a set of Positive Met edges for each species
409 (except *A. thaliana*) based on PlantCyc[37] enzyme annotations, analogous to the Positive Met
410 edges used for *A. thaliana* (Figure 1b). TEA-GCNs consistently outperform conventional non-
411 ensemble PCC-GCNs across all tested species (Figure S8a). Notably, the PCC-GCNs for *A.*
412 *lyrata* and *A. halleri* perform poorly, approaching random classification where AUROC and
413 AUPRC are 0.5, while the TEA-GCNs for the same species perform substantially better. The TEA-
414 GCNs, by contrast, show more consistent and robust performance across all three Arabidopsis
415 species, underscoring their reliability even in data-scarce settings (Figure S8a).

416 To systematically assess cross-species comparability, we developed two complementary
417 conservation metrics based on orthologous edges (see methods): (1) Co-expression correlation,
418 which measures the rank-based similarity in edge weights between GCNs via Spearman's rank
419 Correlation Coefficient (SCC), and (2) Co-expression overlap, which quantifies the agreement in
420 highly co-expressed edges, based on Z(Co-exp. str.) > 1.5, using the Jaccard Index. Applying
421 these metrics to all pairwise combinations of the 10 species (45 comparisons total), we find that
422 TEA-GCNs exhibit markedly higher conservation than PCC-GCNs across both metrics (Figure
423 7b). This suggests that TEA-GCNs not only recover biologically relevant relationships more
424 effectively within species but also yield networks that are more consistent across species
425 boundaries—an essential feature for evolutionary and comparative analyses. To exemplify, the
426 conservation metrics between *A. halleri* (28,722 genes) and *A. thaliana* (27,562 genes)—based
427 on nearly 200 million orthologous edges from 19,989 identified 1-to-1 orthologs (Figure S9)
428 dramatically improved under TEA-GCNs. Specifically, the Co-expression correlation increased
429 from 0.415 (PCC-GCN) to 0.500 (TEA-GCN), and the Co-expression overlap doubled from 0.178
430 to 0.354 (Figure 7b). In other words, the number of conserved, highly co-expressed orthologous
431 edges between these two species doubled when using TEA-GCNs.

432 Overlaying the GCNs of single-orthologs across species allowed us to build a single-
433 ortholog co-expression conservation network of orthologous edges (Figure 7b), where edge co-
434 expression conservation scores are used as edge weights. Briefly, the edge co-expression score
435 for a given orthologous edge is calculated from the number of species it is present in (Species
436 Present) and the number of species it is observed as co-expressed (Species Co-expressed) such
437 that higher scores will be assigned to orthologous edges that are both ubiquitously present and

438 co–expressed in multiple species while orthologous edges that are ubiquitously present yet
439 scarcely co-expressed will be assigned lower scores. We observed that the co-expression
440 conservation network produced by TEA-GCNs across species exhibits higher edge conservation
441 scores than the conservation network produced by PCC GCNs, as TEA-GCN gene co-expression
442 is more frequently preserved across species (Figure S8b and S8C), which, once again,
443 demonstrates its enhanced cross-species comparability. Specifically, 7.44% of the ~ 1.3 million
444 orthologous edges that are present in at least half (≥ 5) of the species set are co-expressed in at
445 least half of the species set according to TEA-GCNs as compared to a meagre 0.27% according
446 to PCC GCNs (Figure S8b and S8c), representing a more than 27-fold difference between the
447 two.

448 To validate the biological significance of co-expression conservation observed across the
449 TEA-GCNs, we calculated the co-expression conservation for single-orthologs in the conservation
450 network (Figure 7c) using a network-theoretical approach via the use of weighted degree
451 centrality, a metric that quantifies the number of edges connecting a given node, such that higher
452 scores will be assigned to single-orthologs with many connections that are highly conserved in
453 co-expression. The conservation scores of single-orthologs were then tested for functional
454 enrichment via GSEA using GO annotations inferred from orthologs in *A. thaliana*. Special care
455 to reduce confirmation bias was taken to not consider GO annotations with “IEP” (Inferred from
456 Expression Patterns). Core cellular processes, including DNA repair, RNA splicing, chromatin and
457 vesicle organisation, cell division, and chloroplast fission (Table 1), were detected to be the most
458 enriched, consistent with conserved housekeeping and organelle-division modules. For example,
459 chloroplast fission genes form a tightly connected, highly conserved neighbourhood in the TEA-
460 GCN conservation network, whereas the corresponding PCC network is much sparser (Figure
461 7d). These results collectively demonstrate that TEA-GCNs yield more biologically meaningful
462 and evolutionarily conserved gene co-expression, positioning them as a powerful framework for
463 cross-species transcriptomic analyses.

464 # Discussion

465 Gene co-expression networks (GCNs) generated from transcriptomic datasets can elucidate the
466 intricate co-regulatory and co-functional relationships between genes, thus enabling gene
467 function prediction. Consequently, GCNs have proven invaluable as a hypothesis generation tool
468 in non-model species, where gene functions are hard to predict based on sequence similarity to
469 characterized homologs in well-studied models (i.e., members of large gene families, orphan and
470 lineage-specific genes/proteins)[76–80]. Due to the increased affordability of next-generation
471 sequencing (NGS), the number of species with transcriptomic data is growing rapidly, which
472 presents an unprecedented opportunity to unravel the evolutionary dynamics of gene function
473 across diverse organisms via the construction of GCNs. However, current GCN construction
474 methodologies[3,26,81,82] were developed for use on small in-house sequenced datasets or manually
475 curated public RNA-seq samples and, therefore, are not well equipped to handle (nor to exploit)
476 the various complexities of public transcriptome data (e.g., high heterogeneity, sample imbalance,
477 batch effects, poor sample quality). To address this unmet need, we developed TEA-GCN, an
478 ensemble methodology centered around determining gene co-expression based on local co-
479 expression observed in different dataset partitions (Figure 1).

480 TEA-GCN is superior to the widely-used method of calculating PCC across the whole
481 dataset (PCC GCN)[26,81,83,84], and to the state-of-the-art Subagging method from the ATTED-II
482 and COXPRESdb databases[15,18] in identifying transregulatory relationships as well as gene co-
483 involvement in enzymatic pathways and biological processes, across three model organisms of
484 vastly different organizational/biological complexities and evolutionary lineages (Figures 2 and
485 S3). In addition, TEA-GCNs possess enhanced across-species comparability and stability
486 compared to PCC GCNs, as seen from their higher conservation of network features across ten
487 Angiosperm species (Figure 7), which can better facilitate kingdom-wide evolutionary studies
488 across GCNs from different species. Evident from the characterization of co-expression edges,
489 TEA-GCN excels in capturing co-expression relationships that manifest in certain conditions
490 and/or in specific tissues, which enables the discovery of complex regulatory mechanisms that
491 can be elusive (Figures 3 and 4). In addition, known functions of seven condition/tissue-specific
492 transcription factor genes (i.e., *WRKY25*, *OBP2*, *ANAC091*, *JLO*, *FUS3*, *WRKY6*, and *ABI5*) can
493 be predicted in their co-expression neighborhoods in TEA-GCN but not in the ensemble GCN
494 from ATTED-II (Figures 4 and S4), exemplifying TEA-GCN's enhanced ability to uncover gene
495 function elusive to traditional co-expression. Furthermore, the targets of TOR kinase, identified

496 via protein-proximity labeling in *A. thaliana*, in another study[85] are strongly corroborated by our *A.*
497 *thaliana* TEA-GCN. Inspired by the growing field of Explainable Artificial Intelligence[86] (XAI)
498 toward explaining the inner workings of complex machine-learning models, TEA-GCN can be
499 made to explain experimental contexts underpinning gene co-expression via NLP. This was
500 possible by capitalizing on our method's ability to capture condition/tissue-specific co-expression.
501 Thus, TEA-GCN can not only outperform more complex state-of-the-art methods but also report
502 biologically significant information as experimental contexts to contextualize gene co-expression.

503 TEA-GCN is designed to operate on complex and heterogeneous public transcriptomic
504 datasets where traditional co-expression inference using PCC is suboptimal. Similar complexity
505 is observed in single-cell RNA-sequencing (scRNA-seq) datasets, which comprise tens of
506 thousands of transcriptomes from diverse cell types. In these datasets, gene expression
507 measurements are often sparse due to limited sequencing depth per cell, which hampers accurate
508 co-expression inference using PCC[87]. Therefore, there is significant potential for extending TEA-
509 GCN to infer gene co-expression networks (GCNs) from scRNA-seq data, highlighting a key
510 direction for future development. Although the ensemble methodology of TEA-GCN vastly
511 improves the quality of GCNs, various aspects of TEA-GCN can still be further optimized. For
512 example, only 3 clustering algorithms – K-means clustering, Hierarchical agglomerative
513 clustering[32], and Density-based spatial clustering of applications with noise (DBSCAN)[33] – were
514 evaluated in this study during TEA-GCN optimization (supplemental results, Figures S3 and S4).
515 A systematic evaluation of the other, more sophisticated or even purpose-built clustering
516 algorithms can help identify ways to generate partitions that can lead to better downstream TEA-
517 GCN performance. In addition, a straightforward method was employed in the study to annotate
518 dataset partitions based on statistically overrepresented lemmas in RNA-seq sample metadata
519 when determining experimental contexts underpinning TEA-GCN ensemble co-expression. In the
520 face of rapid advancements in generative LLMs[88,89], one can tap into LLMs to annotate dataset
521 partitions with greater accuracy and beyond single-word lemmas, when operational costs become
522 less prohibitive. Nonetheless, the robust compatibility of TEA-GCN to work on highly variable
523 datasets without the need for sample annotations for batch correction and/or sample balancing
524 cannot be understated, as high-quality TEA-GCNs can be generated for any species with publicly
525 available RNA-seq samples. This opens the door for scientists working on these species to utilize
526 co-expression to accelerate their research, as well as enabling analyses to be expanded from
527 well-studied models into other species in the same evolutionary group to uncover the evolutionary
528 origins of genes and pathways. The lower barrier to generating quality GCNs using our method

529 can also facilitate the generation of multi-species co-expression networks for downstream
530 analyses. For instance, one can construct genus/clade-specific GCNs (e.g., GCN of cruciferous
531 vegetables) or even GCNs for the Human Pangenome[90]. To take this even further, purpose-built
532 algorithms can be developed to reconstruct ancestral gene co-expression modules from a
533 compendium of extant TEA-GCNs, similar to how ancestral genomes can be inferred from extant
534 genomes[91].

535 Our GRN benchmark is based on rectified correlations and therefore effectively captures
536 only activating / positively co-expressed TF–target relationships, while negatively correlated
537 (putatively repressive) edges are not evaluated. Consequently, our GRN performance estimates
538 should be interpreted as applying primarily to activation, and future extensions of TEA-GCN that
539 preserve and exploit signed correlations will be needed to systematically assess repressive and
540 dual-mode transcriptional regulation. Furthermore, our condition-enrichment explainability layer
541 necessarily relies on the quality of ENA sample metadata; despite the NLP harmonisation
542 (lemmatisation, synonym/ontology mapping, and statistical filtering), these annotations remain
543 heterogeneous and incomplete. As a result, the reported condition enrichments should be viewed
544 as high-precision but lower-recall explanations that are biased towards well-annotated
545 experiments, and they likely miss relevant conditions for samples with sparse or ambiguous
546 metadata. Our NLP-based lemma annotations currently serve as a descriptive aid rather than a
547 fully validated information source. While the examples illustrate that the lemmas often capture the
548 relevant experimental context (Figure 6), we have not yet performed a systematic benchmark
549 against manually curated experiment types or quantified how lemma-based filtering impacts TF
550 function or pathway predictions. A dedicated evaluation of this interpretability layer constitutes an
551 important avenue for future work. It is also important to note that downsampling analyses were
552 not repeated due to the immense computational requirements of generating genome-scale GCN.
553 As a result, the specific stability of TEA-GCN's performance when exactly 5,000 and 500
554 subsamples from *A. thaliana*'s public RNA-seq dataset were used is regrettably unknown.
555 However, the dataset robustness of the TEA-GCN method can adequately be demonstrated from
556 its consistent good performance across the many datasets used in this study that vary greatly in
557 size and composition, and is arguably more indicative for real-world use cases (i.e., Public dataset
558 of 3 model organisms, public dataset of 9 Angiosperms, *A. thaliana* dataset used by ATTED-II
559 and 2 *A. thaliana* subsampling datasets).

560 The TEA-GCN method is available for use via https://github.com/pengkenlim/TEA-GCN
561 as an easy-to-install and well-documented standalone program for users to generate high-quality

562 GCNs from their large transcriptomic datasets. Additionally, the TEA-CGNs of ten Angiosperm
563 species are available in Plant-GCN via https://plantgcn.connectome.tools, a web resource
564 equipped with user-friendly tools for exploration and analysis (Supplementary results, Figures
565 S10 to S12).

566 # Methods

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

569 *S. cerevisiae* and *H. sapiens* mRNA sequences were downloaded from the NCBI RefSeq
570 annotations of their reference genome assemblies (GCF_000146045.2 and GCF_000001405.40).
571 Representative mRNA sequences of protein-coding genes (primary transcripts) for *S. cerevisiae*
572 and *H. sapiens* were then obtained by selecting the longest transcript isoform for every protein-
573 coding gene locus. Primary transcripts of *A. thaliana* were obtained by downloading the
574 "Araport11_cdna_20220914_representative_gene_model.gz" file from the TAIR database[92].
575 6,612 *S. cerevisiae*, 19,379 *H. sapiens*, and 27,562 *A. thaliana* primary transcripts were used as
576 pseudoalignment references for gene expression estimation of RNA-seq sample Illumina
577 sequencing reads (read data) downloaded from ENA using Kallisto's (version 0.50.0) quant
578 functionality in single-end mode with default parameters[93]. Fastq files of 188,189, 80,009, and
579 142,828 RNA-seq samples were downloaded and pseudoaligned for *S. cerevisiae, A. thaliana,*
580 *and H. sapiens,* respectively. Transcripts per million (TPM) values of primary transcripts of RNA-
581 seq samples estimated by Kallisto were used as gene expression abundances to construct gene
582 expression matrices for the different species. RNA-seq samples with poor read alignment were
583 identified based on their %_pseudoaligned statistics (reported by Kallisto) below a certain
584 threshold (5% for *H. sapiens*; 20% for *S. cerevisiae* and *A. thaliana*)[31,94,95]. After the removal of
585 RNA-seq samples with poor read alignment, gene expression matrices comprising gene
586 expression abundances from 150,096 (79.8%), 71,720 (89.6%), and 95,553 (66.9%) retained
587 RNA-seq samples were designated as the public transcriptomic datasets for *S. cerevisiae, A.*
588 *thaliana,* and *H. sapiens,* respectively.

589 To perform batch normalization of the *A. thaliana* RNA-seq dataset containing 14,714
590 accessions found in ATTED-II's dataset, we used the pyComBat[96] python package, which
591 contains the python implementation of the popular ComBat-seq batch correction algorithm[63] used
592 for pre-processing of RNA-seq data by ATTED-II and COXPRESSdb to generate their Subagging
593 GCNs. Similar to ATTED-II and COXPRESSdb, we used bioproject IDs of RNA-seq samples as
594 batches during batch correction using the largest batch as the reference batch[15,18].

595 **Data sources for establishing positive edges**

596 Experimentally verified relationships (positive edges) for Met, TF, and GO categories were
597 established for the three model species (Supplementary Data 8). *S. cerevisiae* TF positive edges
598 are generated by downloading activation regulatory links from the YEASTRACT+ database[97]
599 (http://www.yeastract.com/) that fulfill both "Documented" and "Potential" criteria. According to the
600 authors, "Documented" regulatory links are inferred from experimental transcription factor-binding
601 evidence, while "Potential" regulatory links connect transcription factors with target genes that
602 contain corresponding consensus sequences in the upstream promoter region[97]. *A. thaliana* TF
603 positive edges are generated by downloading activation experimentally validated regulatory links
604 from the PlantTFDB database[98] (https://planttfdb.gao-lab.org/). *H. sapiens* TF positive edges are
605 generated by downloading activation regulatory links from the TRRUST database[99] (version 2;
606 https://www.grnpedia.org/trrust/) and then eliminating regulatory links that do not connect to any
607 known *H. sapiens* transcription factors. *H. sapiens* transcription factors are obtained from the
608 HumanTFs database[100] (https://humantfs.ccbr.utoronto.ca/). Regulatory links in the TRRUST
609 database are considered to be experimentally validated as they were inferred from text-mining of
610 published literature by the authors. GO annotations for S. *cerevisiae*, *A. thaliana,* and *H. sapiens*
611 were downloaded from the EBI Gene Ontology Annotation (GOA) Database[101], The Arabidopsis
612 Information Resource[90] (TAIR), and *Saccharomyces* Genome Database[102] (SGD), respectively.
613 GO positive edges were constructed for each model organism by connecting genes annotated
614 with the same GO BP term, where only experimentally derived annotations with an evidence level
615 of "IDA" were considered. Pathway annotations for S. *cerevisiae*, *A. thaliana,* and *H. sapiens* were
616 webscraped from BioCyc[103], PlantCyc[37], and HumanCyc[104], respectively, where only
617 experimentally derived annotations with an evidence level of "EV-EXP" were considered. Since
618 there is a tendency for co-linearity between the estimated expression abundances of sequence-
619 similar homologs due to how multi-mapping reads are accounted for in pseudoalignment[93], edges
620 that connect genes annotated with the same molecular function were removed from GO positive
621 edges to prevent these edges from inflating performance scores (Figure 1b; blue). Similarly, we
622 removed edges that connect genes coding for enzymes that catalyze the same reaction steps
623 from Met positive edges.

624 Similar to *A. thaliana*, the Positive Met edges of other Angiosperm species in this study
625 were mined from PlantCyc[37]. Since experimentally validated enzyme annotations are scarce or
626 non-existent for the Angiosperm plant species (excluding *A. thaliana*), the definition of these
627 Positive Met edges is broadened for said species to include predicted enzyme-gene associations

while excluding edges involving enzymes predicted to catalyze highly promiscuous reactions (i.e., reactions with > 5 predicted enzymes) to control false positive edges (Supplementary Data 9). These edges were excluded on the assumption that enzymes annotated to catalyze promiscuous reactions are more likely to be inaccurate.

Calculation of AUROC and AUPRC Scores to evaluate GCN performance

Using the co-expression strengths extracted for different sets of positive edges and negative edges (representing genes with no existing functional relationship, obtained from sampling random gene pairs), the Area-under-curve performance scores of the receiver operator characteristic curve (AUROC) and the precision-recall curve (AUPRC) were calculated for each biological aspect (e.g., AUROC-Met and AUROC-Met; Figure 1c). We aggregated the scores with harmonic mean 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. It is important to note that the falsely labelled negative edges (from using random gene pairs with no existing functional relationships) are likely to have a negligible impact on AUROC and AUPRC calculations, even with conservative estimates (Supplemental results).

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:

$$\mu = \frac{\sum_i x_i}{N} \quad (1)$$

$$\sigma = \sqrt{\frac{\sum(X-\mu)}{N}} \quad (2)$$

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

658 $$Z(x_i) = \frac{x_i - \mu}{\sigma} \quad (3)$$

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

661 GRN inference and evaluating GRNs

662 GENIE3[64] GRNs and GRNBoost2[65] GRNs were inferred using the "algo.genie3" and
663 "algo.grnboost2" functions, respectively, that are found in the "arboreto" Python package (version
664 0.1.6). GENIE3 GRNs were inferred using default settings (i.e., ntrees = 1000, tree-method =
665 Random Forest, K = square root method) while GRNBoost2 GRNs were inferred using an
666 early_stop_window_length = 1000. A list of 324 *A. thaliana* transcription factors (Supplementary
667 Data 4) was obtained from PlantTFDB[98] and used as regulators for GENIE3 and GRNBoost2.
668 GRN-inference methods like GENIE3 and GRNBoost2 calculate the propensities of genes as
669 regulatory targets for transcription factors defined by the user. Since GRNs are directional (e.g.,
670 TF1 → TF2) and the aim is to predict TF-target gene pairs (rather than any gene-gene pairs), we
671 defined a new set of positive directional edges from experimentally verified GRNs for *A. thaliana*
672 (Figure 5b; Supplementary Data 4). Negative edges used to calculate AUROC-GRN and AUPRC-
673 GRN were defined by only randomizing target genes connected to edges in the GRN positive
674 edge set (Figure 5b) rather than edges simply connecting two random genes. AUROC-GRN and
675 AUPRC-GRN were calculated 100 times from propensities (for GENIE3 and GRNBoost2 GRNs)
676 and co-expression strengths (for TEA-GCN) of edges (Figure 5b).

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

678 The metadata of RNA-seq samples was downloaded from ENA via the ENA Browser API using
679 "sample_description", "dev_stage", "study_title", "sample_title", "tissue_lib", and "tissue_type" as
680 search fields.

681 Identifying overrepresented lemmas in the *A. thaliana* dataset partitions

682 The "en_core_web_sm" NLP model (version 3.7.1) from the SpaCy (version 3.7.1) Python
683 package was used to lemmatize metadata of *A. thaliana* RNA-seq samples. Before lemmatization,
684 the metadata of RNA-seq samples was preprocessed. First, metadata returned from all search

fields of every RNA-seq sample were concatenated. Next, stop words (e.g., about, quite, the) were detected from the concatenated metadata using the “token.is_stop” attribute of word tokens generated by the “en_core_web_sm” model. Digits and punctuation were also removed from the concatenated metadata.

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

Let $S = [s_1, s_2, s_3, \ldots s_n]$ be a vector of permuted occurrences of a lemma in a dataset partition from n random shufflings of sample-partition assignments, and o be the observed occurrence. The formula is the standard empirical permutation p-value used in randomisation tests and Monte-Carlo hypothesis testing[105]. Specifically, we compute:

$$P = \frac{1 + \sum_{i=1}^{n} I(s_i \geq o)}{1+n} \quad (4)$$

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[106], 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 (Supplementary Data 6).

Lemma annotations were used purely as an interpretability layer and did not influence TEA-GCN training or evaluation. For each TF–target edge, we extracted sentences in which both genes co-occurred and stored lemmatized tokens that describe the experimental context (e.g., “overexpression line”, “knockout”, “root”, “drought”). These lemmas are exposed in the web interface to help users understand under which experimental conditions the interaction was reported.

716 **Enrichment-based analyses**

717 GSEA of co-expression neighborhoods of selected *A. thaliana* transcription factors was
718 implemented by the GSEApy python package[107] (version 1.1.2). For each transcription factor,
719 genes were ranked based on co-expression strength. The ranks were centered, and gene sets
720 were constructed using *A. thaliana* GO annotations (biological process category). Together, the
721 centered ranks and constructed gene sets were used as input for the prerank sub-command of
722 GSEApy (weight = 1). Annotating edges between co-expressed genes with experimental context
723 lemmas also uses the prerank sub-command of GSEApy in the same way. However, centered
724 partition rankings determined from ranking dataset partitions based on co-expression strengths
725 after Coefficient Aggregation were used in place of gene rankings. Instead of gene sets, dataset
726 partitions for which lemma labels were assigned were used.

727 **Calculating GCN conservation between species and building single-ortholog co-**
728 **expression conservation networks**

729 Orthologous edges connecting shared 1-to-1 orthologs can be identified between the GCNs of
730 any two species (Figure 8c). Pairwise conservation between GCNs of any two species can be
731 inferred using Co-expression correlation and Co-expression Overlap. The former calculates the
732 Spearman's rank Correlation Coefficient[108] of co-expression strengths of shared orthologous
733 edges between the two GCNs. On the other hand, Co-expression Overlap can be calculated
734 based on the Jaccard Index. The mathematical description for calculating Co-expression Overlap
735 is detailed below:

736 $$Co\text{-}expression\ Overlap = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

737 Where A and B are sets of orthologous edges in the Ortholog GCNs considered to be co-
738 expressed according to the GCN of the first species and second species, respectively, using a
739 co-expression threshold of Z(co-exp.) ≥ 1.5. Orthology between Angiosperm genes (Figure S9)
740 was inferred from OrthoFinder2[109] based on their primary transcript sequences (Supplemental
741 methods).

742 Single-ortholog co-expression conservation networks were built from GCNs using
743 orthologous edges that were shared between at least 2 species. Edge Co-expression
744 Conservation (ECC, ECC) score for each orthologous edge can be calculated using No. of Species
745 Present (SP, SP) and Species Co-expressed (SC, SC), which represents the number of species

746 for which the orthologous edge connects between two single-orthologs and the number of species
747 in Species Present for which the orthologous edge is considered co-expressed (Z(co-exp.) ≥ 1.5),
748 respectively:

749 $$ECC = \frac{SC^2}{SP} \quad (6)$$

750 Since high SP orthologous edges are, by nature of being more ubiquitous across species,
751 separated by a longer evolutionary distance, we think that significant co-expression conservation
752 can be indicated with a proportionally lower SC as SP of orthologous edges increases. As such,
753 ECC intentionally prioritizes SC over SP by being second-order sensitive to SC but only inversely
754 linear to SP.

755 Single-Ortholog Conservation Scores for single-orthologs in the Single-ortholog co-
756 expression conservation network were inferred using the weighted degree centrality[110].

757 Data Availability

758 The model organism TEA-GCN data generated in this study have been deposited in the FigShare
759 database under accession identifiers 0.6084/m9.figshare.26347018
760 [https://doi.org/10.6084/m9.figshare.26347018], 10.6084/m9.figshare.26347078
761 [https://doi.org/10.6084/m9.figshare.26347078], and 10.6084/m9.figshare.26347297
762 [https://doi.org/10.6084/m9.figshare.26347297]. Transcriptomic datasets in the form of gene
763 expression matrices used in this study can be found under 10.6084/m9.figshare.26346916
764 [https://doi.org/10.6084/m9.figshare.26346916], 10.6084/m9.figshare.26346964
765 [https://doi.org/10.6084/m9.figshare.26346964], and 10.6084/m9.figshare.26347015
766 [https://doi.org/10.6084/m9.figshare.26347015]. Other data generated in this study are provided
767 in the Supplementary Data and Source Data files. TEA-GCNs for 10 Angiosperms can be
768 accessed via the Plant-GCN database at https://plantgcn.connectome.tools/. Source Data is
769 provided with this paper.

770 Code Availability

771 TEA-GCN is available at https://github.com/pengkenlim/TEA-GCN[111].

772 # References

773 1. Usadel, B. *et al.* Co-expression tools for plant biology: opportunities for hypothesis
774 generation and caveats. *Plant Cell Environ.* **32**, 1633–1651 (2009).

775 2. Lim, P. K., Zheng, X., Goh, J. C. & Mutwil, M. Exploiting plant transcriptomic databases:
776 Resources, tools, and approaches. *Plant Commun.* **3**, 100323 (2022).

777 3. Rao, X. & Dixon, R. A. Co-expression networks for plant biology: why and how. *Acta*
778 *Biochim. Biophys. Sin.* **51**, 981–988 (2019).

779 4. Aoki, K., Ogata, Y. & Shibata, D. Approaches for extracting practical information from gene
780 co-expression networks in plant biology. *Plant Cell Physiol.* **48**, 381–390 (2007).

781 5. Rhee, S. Y. & Mutwil, M. Towards revealing the functions of all genes in plants. *Trends*
782 *Plant Sci.* **19**, 212–221 (2014).

783 6. Julca, I., Tan, Q. W. & Mutwil, M. Toward kingdom-wide analyses of gene expression.
784 *Trends Plant Sci.* **28**, 235–249 (2023).

785 7. Burgin, J. *et al.* The European Nucleotide Archive in 2022. *Nucleic Acids Res.* **51**, D121–
786 D125 (2023).

787 8. Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple
788 species. *Nucleic Acids Res.* **50**, D129–D140 (2022).

789 9. Ruprecht, C., Vaid, N., Proost, S., Persson, S. & Mutwil, M. Beyond Genomics: Studying
790 Evolution with Gene Coexpression Networks. *Trends Plant Sci.* **22**, 298–307 (2017).

791 10. Ovens, K., Eames, B. F. & McQuillan, I. Comparative Analyses of Gene Co-expression
792 Networks: Implementations and Applications in the Study of Evolution. *Front. Genet.* **12**, 695399
793 (2021).

794 11. Vandenbon, A. Evaluation of critical data processing steps for reliable prediction of gene
795 co-expression from large collections of RNA-seq data. *PLOS ONE* **17**, e0263344 (2022).

796 12. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network
797 construction and analysis: safety in numbers. *Bioinforma. Oxf. Engl.* **31**, 2123–2130 (2015).

798 13. Gupta, C. & Pereira, A. Recent advances in gene function prediction using context-
799 specific coexpression networks in plants. *F1000Research* **8**, F1000 Faculty Rev-153 (2019).

800 14. Pearce, S., Ferguson, A., King, J. & Wilson, Z. A. FlowerNet: a gene expression
801 correlation network for anther and pollen development. *Plant Physiol.* **167**, 1717–1730 (2015).

802 15. Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an
803 animal gene coexpression database navigating from a global view to detailed investigations.
804 *Nucleic Acids Res.* **51**, D80–D87 (2023).

805 16. Lin, H., Yu, J., Pearce, S. P., Zhang, D. & Wilson, Z. A. RiceAntherNet: a gene co-
806 expression network for identifying anther and pollen development genes. *Plant J. Cell Mol. Biol.*
807 **92**, 1076–1091 (2017).

808 17. Lancour, D. *et al.* Analysis of brain region-specific co-expression networks reveals
809 clustering of established and novel genes associated with Alzheimer disease. *Alzheimers Res.*
810 *Ther.* **12**, 103 (2020).

811 18. Obayashi, T., Hibara, H., Kagaya, Y., Aoki, Y. & Kinoshita, K. ATTED-II v11: A Plant
812 Gene Coexpression Database Using a Sample Balancing Technique by Subagging of Principal
813 Components. *Plant Cell Physiol.* **63**, 869–881 (2022).

814 19. Liesecke, F. *et al.* Improved gene co-expression network quality through expression
815 dataset down-sampling and network aggregation. *Sci. Rep.* **9**, 14431 (2019).

816 20. Wong, D. C. J. Network aggregation improves gene function prediction of grapevine
817 gene co-expression networks. *Plant Mol. Biol.* **103**, 425–441 (2020).

818 21. Arshad, Z. & McDonald, J. F. A computational approach to generate highly conserved
819 gene co-expression networks with RNA-seq data. *STAR Protoc.* **3**, 101432 (2022).

820 22. Liesecke, F. *et al.* Ranking genome-wide correlation measurements improves microarray
821 and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* **8**, 10885 (2018).

822 23. Hou, J. *et al.* Distance correlation application to gene co-expression network analysis.
823 *BMC Bioinformatics* **23**, 81 (2022).

824 24. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual
825 information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328 (2012).

826 25. Hardin, J., Mitani, A., Hicks, L. & VanKoten, B. A robust measure of correlation between
827 two genes on a microarray. *BMC Bioinformatics* **8**, 220 (2007).

828 26. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
829 analysis. *BMC Bioinformatics* **9**, 559 (2008).

830 27. Julca, I. *et al.* Comparative transcriptomic analysis reveals conserved programmes
831 underpinning organogenesis and reproduction in land plants. *Nat. Plants* **7**, 1143–1159 (2021).

832 28. Rosenberg, A. & Hirschberg, J. V-Measure: A Conditional Entropy-Based External
833 Cluster Evaluation Measure. in *Proceedings of the 2007 Joint Conference on Empirical Methods*
834 *in Natural Language Processing and Computational Natural Language Learning (EMNLP-*
835 *CoNLL)* (ed. Eisner, J.) 410–420 (Association for Computational Linguistics, Prague, Czech
836 Republic, 2007).

837 29. Arthur, D. & Vassilvitskii, S. k-means++: the advantages of careful seeding. in
838 *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–
839 1035 (Society for Industrial and Applied Mathematics, USA, 2007).

840 30. Zheng, X., Lim, P. K., Mutwil, M. & Wang, Y. A method for mining condition-specific co-
841 expressed genes in Camellia sinensis based on k-means clustering. *BMC Plant Biol.* **24**, 373
842 (2024).

843 31. Lim, P. K., Wang, R. & Mutwil, M. LSTrAP-denovo: Automated Generation of
844 Transcriptome Atlases for Eukaryotic Species Without Genomes. *Physiol. Plant.* **176**, e14407
845 (2024).

846 32. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at
847 https://doi.org/10.48550/arXiv.1109.2378 (2011).

848 33. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering
849 clusters in large spatial databases with noise. in *Proceedings of the Second International*

850 *Conference on Knowledge Discovery and Data Mining* 226–231 (AAAI Press, Portland, Oregon,
851 1996).

852 34. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of
853 cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

854 35. Aravind, L., Anantharaman, V. & Venancio, T. M. Apprehending multicellularity:
855 regulatory networks, genomics, and evolution. *Birth Defects Res. Part C Embryo Today Rev.*
856 **87**, 143–164 (2009).

857 36. Anantharaman, V., Iyer, L. M. & Aravind, L. Comparative genomics of protists: new
858 insights into the evolution of eukaryotic signal transduction and gene regulation. *Annu. Rev.*
859 *Microbiol.* **61**, 453–475 (2007).

860 37. Hawkins, C. *et al.* Plant Metabolic Network 15: A resource of genome-wide metabolism
861 databases for 126 plants and algae. *J. Integr. Plant Biol.* **63**, 1888–1905 (2021).

862 38. Delli-Ponti, R., Shivhare, D. & Mutwil, M. Using Gene Expression to Study Specialized
863 Metabolism-A Practical Guide. *Front. Plant Sci.* **11**, 625035 (2020).

864 39. Brockington, S. F. *et al.* Lineage-specific gene radiations underlie the evolution of novel
865 betalain pigmentation in Caryophyllales. *New Phytol.* **207**, 1170–1180 (2015).

866 40. Wink, M. Evolution of secondary metabolites from an ecological and molecular
867 phylogenetic perspective. *Phytochemistry* **64**, 3–19 (2003).

868 41. Lim, P. K., Julca, I. & Mutwil, M. Redesigning plant specialized metabolism with
869 supervised machine learning using publicly available reactome data. *Comput. Struct.*
870 *Biotechnol. J.* **21**, 1639–1650 (2023).

871 42. Colinas, M. & Goossens, A. Combinatorial Transcriptional Control of Plant Specialized
872 Metabolism. *Trends Plant Sci.* **23**, 324–336 (2018).

873 43. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends*
874 *Plant Sci.* **24**, 83–98 (2019).

875 44. Voß, U., Bishopp, A., Farcot, E. & Bennett, M. J. Modelling hormonal response and
876 development. *Trends Plant Sci.* **19**, 311–319 (2014).

877 45. Zhang, Y., Berman, A. & Shani, E. Plant Hormone Transport and Localization: Signaling
878 Molecules on the Move. *Annu. Rev. Plant Biol.* **74**, 453–479 (2023).

879 46. Verma, V., Ravindran, P. & Kumar, P. P. Plant hormone-mediated regulation of stress
880 responses. *BMC Plant Biol.* **16**, 86 (2016).

881 47. Prieto, M. A., López, C. J. & Simal-Gandara, J. Glucosinolates: Molecular structure,
882 breakdown, genetic, bioavailability, properties and healthy and adverse effects. *Adv. Food Nutr.*
883 *Res.* **90**, 305–350 (2019).

884 48. Barba, F. J. *et al.* Bioavailability of Glucosinolates and Their Breakdown Products:
885 Impact of Processing. *Front. Nutr.* **3**, 24 (2016).

886 49. Johnson, I. T. Glucosinolates: bioavailability and importance to health. *Int. J. Vitam. Nutr.*
887 *Res. Int. Z. Vitam.- Ernahrungsforschung J. Int. Vitaminol. Nutr.* **72**, 26–31 (2002).

888 50. Sønderby, I. E., Geu-Flores, F. & Halkier, B. A. Biosynthesis of glucosinolates--gene
889 discovery and beyond. *Trends Plant Sci.* **15**, 283–290 (2010).

890 51. Pfalz, M. *et al.* Metabolic engineering in Nicotiana benthamiana reveals key enzyme
891 functions in Arabidopsis indole glucosinolate modification. *Plant Cell* **23**, 716–729 (2011).

892 52. Bureau, M. & Simon, R. JLO regulates embryo patterning and organ initiation by
893 controlling auxin transport. *Plant Signal. Behav.* **3**, 145–147 (2008).

894 53. Collin, A., Daszkowska-Golec, A. & Szarejko, I. Updates on the Role of ABSCISIC ACID
895 INSENSITIVE 5 (ABI5) and ABSCISIC ACID-RESPONSIVE ELEMENT BINDING FACTORs
896 (ABFs) in ABA Signaling in Different Developmental Stages in Plants. *Cells* **10**, 1996 (2021).

897 54. Moreno-Risueno, M. A. *et al.* FUSCA3 from barley unveils a common transcriptional
898 regulation of seed-specific genes between cereals and Arabidopsis. *Plant J. Cell Mol. Biol.* **53**,
899 882–894 (2008).

900 55. Skirycz, A. *et al.* DOF transcription factor AtDof1.1 (OBP2) is part of a regulatory
901 network controlling glucosinolate biosynthesis in Arabidopsis. *Plant J. Cell Mol. Biol.* **47**, 10–24
902 (2006).

903 56. Podzimska-Sroka, D., O'Shea, C., Gregersen, P. L. & Skriver, K. NAC Transcription
904 Factors in Senescence: From Molecular Structure to Function in Crops. *Plants Basel Switz.* **4**,
905 412–448 (2015).

906 57. Chen, Y.-F. *et al.* The WRKY6 transcription factor modulates PHOSPHATE1 expression
907 in response to low Pi stress in Arabidopsis. *Plant Cell* **21**, 3554–3566 (2009).

908 58. Jiang, Y. & Deyholos, M. K. Functional characterization of Arabidopsis NaCl-inducible
909 WRKY25 and WRKY33 transcription factors in abiotic stresses. *Plant Mol. Biol.* **69**, 91–105
910 (2009).

911 59. Doll, J. *et al.* Arabidopsis thaliana WRKY25 Transcription Factor Mediates Oxidative
912 Stress Tolerance and Regulates Senescence in a Redox-Dependent Manner. *Front. Plant Sci.*
913 **10**, 1734 (2019).

914 60. Li, S., Fu, Q., Chen, L., Huang, W. & Yu, D. Arabidopsis thaliana WRKY25, WRKY26,
915 and WRKY33 coordinate induction of plant thermotolerance. *Planta* **233**, 1237–1252 (2011).

916 61. Li, S., Fu, Q., Huang, W. & Yu, D. Functional analysis of an Arabidopsis transcription
917 factor WRKY25 in heat stress. *Plant Cell Rep.* **28**, 683–693 (2009).

918 62. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
919 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

920 63. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for
921 RNA-seq count data. *NAR Genomics Bioinforma.* **2**, lqaa078 (2020).

922 64. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks
923 from expression data using tree-based methods. *PloS One* **5**, e12776 (2010).

924 65. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene
925 regulatory networks. *Bioinforma. Oxf. Engl.* **35**, 2159–2161 (2019).

926 66. Kavi Kishor, P. B., Tiozon, R. N., Fernie, A. R. & Sreenivasulu, N. Abscisic acid and its
927 role in the modulation of plant growth, development, and yield stability. *Trends Plant Sci.* **27**,
928 1283–1295 (2022).

929 67. Bulgakov, V. P., Wu, H.-C. & Jinn, T.-L. Coordination of ABA and Chaperone Signaling
930 in Plant Stress Responses. *Trends Plant Sci.* **24**, 636–651 (2019).

931 68. Nambara, E. & Marion-Poll, A. ABA action and interactions in seeds. *Trends Plant Sci.* **8**,
932 213–217 (2003).

933 69. Lopez-Molina, L., Mongrand, S., McLachlin, D. T., Chait, B. T. & Chua, N.-H. ABI5 acts
934 downstream of ABI3 to execute an ABA-dependent growth arrest during germination. *Plant J.*
935 *Cell Mol. Biol.* **32**, 317–328 (2002).

936 70. Pollo-Rodríguez, F., Sánchez-Vicente, I. & Lorenzo, O. The turnover of ABI5 by scaffold
937 proteins to attenuate ABA signaling. *J. Exp. Bot.* **75**, 3749–3753 (2024).

938 71. Peng, J. *et al.* COP1 positively regulates ABA signaling during Arabidopsis seedling
939 growth in darkness by mediating ABA-induced ABI5 accumulation. *Plant Cell* **34**, 2286–2308
940 (2022).

941 72. Qi, L. *et al.* PHYTOCHROME-INTERACTING FACTORS Interact with the ABA
942 Receptors PYL8 and PYL9 to Orchestrate ABA Signaling in Darkness. *Mol. Plant* **13**, 414–430
943 (2020).

944 73. Xiong, S.-X. *et al.* The transcription factors MS188 and AMS form a complex to activate
945 the expression of CYP703A2 for sporopollenin biosynthesis in Arabidopsis thaliana. *Plant J. Cell*
946 *Mol. Biol.* **88**, 936–946 (2016).

947 74. Wang, K. *et al.* The Regulation of Sporopollenin Biosynthesis Genes for Rapid Pollen
948 Wall Formation. *Plant Physiol.* **178**, 283–294 (2018).

949 75. Zhang, Z.-B. *et al.* Transcription factor AtMYB103 is required for anther development by
950 regulating tapetum development, callose dissolution and exine formation in Arabidopsis. *Plant J.*
951 *Cell Mol. Biol.* **52**, 528–538 (2007).

952 76. Gerlt, J. A., Babbitt, P. C., Jacobson, M. P. & Almo, S. C. Divergent evolution in enolase
953 superfamily: strategies for assigning functions. *J. Biol. Chem.* **287**, 29–34 (2012).

954 77. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends*
955 *Plant Sci.* **19**, 698–708 (2014).

956 78. Jeffery, C. J. Current successes and remaining challenges in protein function prediction.
957 *Front. Bioinforma.* **3**, 1222182 (2023).

958 79. Jeffery, C. J. Enzymes, pseudoenzymes, and moonlighting proteins: diversity of function
959 in protein superfamilies. *FEBS J.* **287**, 4141–4149 (2020).

960 80. Jacobson, M. P., Kalyanaraman, C., Zhao, S. & Tian, B. Leveraging structure for
961 enzyme function prediction: methods, opportunities, and challenges. *Trends Biochem. Sci.* **39**,
962 363–371 (2014).

963 81. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and
964 Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2015).

965 82. Zainal-Abidin, R.-A., Harun, S., Vengatharajuloo, V., Tamizi, A.-A. & Samsulrizal, N. H.
966 Gene Co-Expression Network Tools and Databases for Crop Improvement. *Plants Basel Switz.*
967 **11**, 1625 (2022).

968 83. Proost, S. & Mutwil, M. CoNekT: an open-source framework for comparative genomic
969 and transcriptomic network analyses. *Nucleic Acids Res.* **46**, W133–W140 (2018).

970 84. De Bodt, S., Hollunder, J., Nelissen, H., Meulemeester, N. & Inzé, D. CORNET 2.0:
971 integrating plant coexpression, protein-protein interactions, regulatory interactions, gene
972 associations and functional annotations. *New Phytol.* **195**, 707–720 (2012).

973 85. Zheng, S. *et al.* Pupylation-Based Proximity Labeling Unravels a Comprehensive Protein
974 and Phosphoprotein Interactome of the Arabidopsis TOR Complex. *Adv. Sci. Weinh. Baden-*
975 *Wurtt. Ger.* **12**, e2414496 (2025).

976 86. A., S. & R., S. A systematic review of Explainable Artificial Intelligence models and
977 applications: Recent developments and future trends. *Decis. Anal. J.* **7**, 100230 (2023).

978 87. Vivian Li, W. & Li, Y. scLink: Inferring Sparse Gene Co-expression Networks from
979 Single-cell Expression Data. *Genomics Proteomics Bioinformatics* **19**, 475–492 (2021).

980 88. Brown, T. *et al.* Language Models are Few-Shot Learners. in *Advances in Neural*
981 *Information Processing Systems* vol. 33 1877–1901 (Curran Associates, Inc., 2020).

982 89. Guo, D. *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement
983 learning. *Nature* **645**, 633–638 (2025).

984 90. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

985 91. Muffato, M. *et al.* Reconstruction of hundreds of reference ancestral genomes across the
986 eukaryotic kingdom. *Nat. Ecol. Evol.* **7**, 355–366 (2023).

987 92. Garcia-Hernandez, M. *et al.* TAIR: a resource for integrated Arabidopsis data. *Funct.*
988 *Integr. Genomics* **2**, 239–253 (2002).

989 93. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
990 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

991 94. Villanueva, E. M. F. *et al.* Protist.guru: A Comparative Transcriptomics Database for
992 Protists. *J. Mol. Biol.* **434**, 167502 (2022).

993 95. Lim, P. K. *et al.* Bacteria.guru: Comparative Transcriptomics and Co-Expression
994 Database for Bacterial Pathogens. *J. Mol. Biol.* **434**, 167380 (2022).

995 96. Behdenna, A. *et al.* pyComBat, a Python tool for batch effects correction in high-
996 throughput molecular data using empirical Bayes methods. *BMC Bioinformatics* **24**, 459 (2023).

997 97. Monteiro, P. T. *et al.* YEASTRACT+: a portal for cross-species comparative genomics of
998 transcription regulation in yeasts. *Nucleic Acids Res.* **48**, D642–D649 (2020).

999 98. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory
1000 interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).

1001 99. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse
1002 transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

1003 100. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

1004 101. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in
1005 Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, D262-266 (2004).

1006 102. Wong, E. D. *et al.* Saccharomyces genome database update: server architecture, pan-
1007 genome nomenclature, and external resources. *Genetics* **224**, iyac191 (2023).

1008 103. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways.
1009 *Brief. Bioinform.* **20**, 1085–1093 (2019).

1010 104. Romero, P. *et al.* Computational prediction of human metabolic pathways from the
1011 complete human genome. *Genome Biol.* **6**, R2 (2005).

1012 105. Good, P. I. *Permutation, Parametric, and Bootstrap Tests of Hypotheses (Springer*
1013 *Series in Statistics)*. (Springer-Verlag, Berlin, Heidelberg, 2004).

1014 106. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
1015 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

1016 107. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set
1017 enrichment analysis in Python. *Bioinforma. Oxf. Engl.* **39**, btac757 (2023).

1018 108. Spearman, C. The proof and measurement of association between two things. *Int. J.*
1019 *Epidemiol.* **39**, 1137–1150 (2010).

1020 109. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1021 genomics. *Genome Biol.* **20**, 238 (2019).

1022 110. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of
1023 complex weighted networks. *Proc. Natl. Acad. Sci.* **101**, 3747–3752 (2004).

1024 111. Lim, P. K., Wang, R., Lim, S. C., Antony Velankanni, J. P., & Mutwil, M. (2026).
1025 Constructing Gene Co-functional and Co-regulatory Networks from Public Transcriptomes using
1026 Condition-Specific Ensemble Co-expression. Zenodo, TEA-GCN (v1.0.0),
1027 https://doi.org/10.5281/zenodo.19143726 (2020).

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 the 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 and 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.

Competing Interests Statement

The authors declare no competing interests.

1062 # Tables

1063 **Table 1. Enriched Gene Ontology (GO) terms of Single-orthologs with high co-expression conservation**

GO Term ID	GO Type[§]	GO Term	NES*	Recall[†]	No. Single-orthologs (Leading / Total)	P-value (BH adjusted)[◊]
GO:0040029	BP	epigenetic regulation of gene expression	1.684978	0.612245	30/49	0.004789
GO:0006281	BP	DNA repair	1.68368	0.529412	36/68	0.004887
GO:0000911	BP	cytokinesis by cell plate formation	1.668684	0.666667	26/39	0.004842
GO:0000724	BP	double-strand break repair via homologous recombination	1.6537	0.545455	18/33	0.007355
GO:0008380	BP	RNA splicing	1.647197	0.560976	23/41	0.007339
GO:0005769	CC	early endosome	1.639101	0.526316	1019	0.008894
GO:0010020	BP	chloroplast fission	1.632415	0.666667	14/21	0.010311
GO:0006897	BP	endocytosis	1.627713	0.692308	9/13	0.012361
GO:0031010	CC	ISWI-type complex	1.625973	0.692308	9/13	0.012786
GO:0035194	BP	regulatory ncRNA-mediated post-transcriptional gene silencing	1.625927	0.545455	6/11	0.012481
GO:0006397	BP	mRNA processing	1.622777	0.678571	19/28	0.013058
GO:0006310	BP	DNA recombination	1.616724	0.722222	13/18	0.013846
GO:0032991	CC	protein-containing complex	1.612973	0.535714	15/28	0.015478
GO:0005802	CC	trans-Golgi network	1.60647	0.509653	132/259	0.016151
GO:0009294	BP	DNA-mediated transformation	1.602963	0.56	14/25	0.016398
GO:0006302	BP	double-strand break repair	1.598286	0.6	9/15	0.017671
GO:0045814	BP	negative regulation of gene expression, epigenetic	1.596284	0.771429	27/35	0.018473
GO:0000785	CC	chromatin	1.596079	0.545455	12/22	0.018313
GO:0048589	BP	developmental growth	1.595159	0.571429	8/14	0.018464
GO:0003924	MF	GTPase activity	1.592495	0.535714	15/28	0.01892

1064 [§]*BP, CC, and MF correspond to the different GO categories of "Biological Process", "Cellular Component", and "Molecular Function",*
1065 *respectively.*
1066 * *In the interest of brevity, only the top 20 GO terms (sorted by NES) are shown.*
1067 [†] *Only highly specific GO terms with a large proportion of Single-orthologs with high co-expression conservation with recall ≥ 0.5 are*
1068 *shown.*
1069 [◊] *In accordance with GSEA, p-values were generated via permutation testing as a fraction of ES from 10,000 permutated gene lists*
1070 *that are larger than the ES observed. p-values were then subsequently subjected to Benjamini-Hochberg correction.*

1071 # Figure Legends

1072 **Figure 1. Overview of the Two-Tier Ensemble Aggregation Gene Co-expression Network**
1073 **(TEA-GCN) framework and GCN evaluation methodology.**

1074 **a** The processing of data from an initial gene expression matrix to the final ensemble GCN is
1075 visualized using a toy model in which there are only three genes (A, B, C) in the genome. The
1076 gene expression matrix is first partitioned via k-means clustering at a clustering granularity (k=3)
1077 that has been determined using a combination of semi-supervision and/or internal clustering
1078 evaluation metrics. For each dataset partition, GCNs are generated based on three different
1079 correlation coefficients and aggregated into one ensemble GCN. Resultant GCNs from each
1080 partition are again aggregated to yield the final ensemble TEA-GCN. GCNs with co-expression
1081 strength represented by Pearson's correlation coefficient (PCC), Spearman's rank correlation
1082 coefficient (SCC), and Biweight midcorrelation (Bicor) are denoted by purple, brown, and blue
1083 edges, respectively. The node color of GCNs corresponds to the particular dataset partition they
1084 represent. **b** We define three sets of positive edges connecting genes sharing different biological
1085 relationships. Positive Metabolic (Met) edges are made by pairing experimentally validated
1086 enzymes from the same pathway. Edges connecting enzymes that catalyze the same reaction
1087 step are removed from positive Met edges. Positive Transcription factor (TF) edges are generated
1088 between genes and their corresponding activator TFs. Positive Gene Ontology (GO) edges are
1089 made from pairing genes that were annotated for the same biological process (BP) GO term at
1090 the IDA (inferred from direct assay) evidence level. Edges connecting genes annotated with the
1091 same molecular function (MF)-GO term are removed from positive GO edges. **b** To evaluate a
1092 given GCN, co-expression strengths of positive edge sets for each biological aspect are
1093 compared to co-expression strengths of randomly determined negative edges of the same size
1094 using the area-under-curve (AUC) of the receiver operator characteristics curve (ROC) and
1095 precision-recall curve (PRC). Scores calculated for each positive-negative edge set pair represent
1096 a GCN's performance to identify different biologically significant relationships between genes and
1097 can be aggregated using the harmonic mean to infer a GCN's overall performance.

1098

1099 **Figure 2. Benchmarking TEA-GCN across three different model organisms against**
1100 **Subagging ensemble GCNs.**

1101 **a** Distribution of Measurements of AUROC-HM performance scores between Subagging CGNs
1102 and TEA-GCNs across *S. cerevisiae*, *A. thaliana,* and *H. sapiens*. **b** Distribution of Measurements
1103 of AUPRC-HM performance scores. Each measurement refers to a score calculated from a set
1104 of randomly generated negative edges. For panels a and b, We report on the delta values of
1105 performance score measurements (i.e., Δ AUROC-HM and Δ AUPRC-HM) used to compare
1106 GCNs on a unified axis. The delta values were calculated by subtracting the performance score
1107 measurements from the median performance score measurement of the baseline non-ensemble
1108 PCC network of respective organisms. Median performance score measurements of PCC (blue)
1109 networks, Subagging CGNs (orange), and TEA-GCNs (green) are indicated. Student's t-test (2-
1110 sided) p-values comparing performance measurements of ≤ 0.05, ≤ 0.01, and ≤ 0.001 were
1111 denoted by *, **, and *** respectively, while the adjacent numbers indicate the difference between
1112 medians was shown together with 95% confidence intervals (Difference [Lower interval, Upper
1113 interval]). In Violin plots, the centre, upper and lower bounds of boxes correspond to median, 75th
1114 and 95th percentiles, while the upper and lower bounds of whiskers correspond to maximum and
1115 minimum values. Exact p-values can be found in source data. **c** Receiver operating characteristic
1116 curves (ROCs) correspond to median performance scores for each biological aspect and
1117 organism. **d** Precision-recall curves (PRCs) correspond to median performance scores for each
1118 biological aspect and organism. Across all panels, performance information relating to the PCC
1119 network, Subagging CGN, and TEA-GCN is represented in blue, orange, and green, respectively.
1120 Source data for all panels are provided in the Source Data file.

1121

1122 **Figure 3. Characterizing TEA-GCN's ability to capture metabolic pathways in *A. thaliana*.**

1123 **a** Proportion of Met positive edges of different PlantCyc pathways that are co-expressed in *A.*
1124 *thaliana* TEA-GCN and ATTED-II GCN at two (>0.5 and >=1) co-expression thresholds. Met
1125 positive edges connect experimentally validated enzymes of different reaction steps. The grey-
1126 labelled plot corresponds to first-level ontological classes and compares major pathway
1127 categories. The blue-labeled plot breaks down second-level ontological classes within
1128 "Biosynthesis", of which third-level ontological classes consisting of "Aromatic Compound
1129 Biosynthesis", "Cell Structure Biosynthesis", "Hormone and Signalling Molecule Biosynthesis",
1130 and "Secondary Metabolite Biosynthesis" were further broken down in the brown, purple, yellow,
1131 and pink plots, respectively. **b** Co-expression between enzymes of "Glucosinolate Biosynthesis
1132 from Tryptophan" PlantCyc pathway (PWY-601) in *A. thaliana* TEA-GCN and ATTED-II GCN.

1133 Nodes represent experimentally validated enzymes according to PlantCyc and are labeled with
1134 gene symbols (if available) or gene identifiers. The thickness of grey edges connecting nodes
1135 corresponds to standardized co-expression strengths between enzymes in the respective GCNs.
1136 **c** Reaction steps in Glucosinolate Biosynthesis from Tryptophan PWY-601. Colored blocks in the
1137 middle of reaction steps annotate enzyme nodes in panel B based on the reaction step in which
1138 they participate. Source data for panels a and b are provided in the Source Data file.

1139

1140 **Figure 4. Characterization of transcription factor-target predictions in *A. thaliana* with TEA-**
1141 **GCN.**

1142 **a** Venn diagram (top) showing 457 common and unique TF positive edges (out of 673 total) that
1143 are co-expressed (Z(Co-exp.) ≥ 1) in *A. thaliana* TEA-GCN and ATTED-II GCN. Heatmap
1144 (bottom) shows co-expression strengths of TF positive edges in the two GCNs. Edges are sorted
1145 based on their co-expression in ATTED-II GCN. Color intensities correspond to co-expression
1146 strength (i.e., Z(co-exp.)). **b** Co-expression between selected TFs and their known targets in
1147 TEA-GCN and ATTED-II GCN. TF and non-TF genes are represented by yellow diamond-shaped
1148 and purple circular-shaped nodes, respectively. Edge thickness represents co-expression
1149 strength between TFs and their known targets. Grey chevrons indicate the direction of regulation
1150 between TFs and their targets. **c** Left: Schematic of enrichment score calculation from co-
1151 expression neighbourhood. Right: Dot plots comparing the enrichment statistics of selected GO
1152 terms in the neighborhood of *WRKY25* in TEA-GCN and in ATTED-II GCN. GO terms were
1153 selected based on the known functions of *WRKY25*. Dot sizes correspond to normalized
1154 enrichment scores (NES), while inverted triangles represent GO terms with negative NES scores.
1155 For GO terms with positive NES scores, the x-axis indicates the statistical significance in the form
1156 of Benjamini-Hochberg-corrected p-values (p_BH). In accordance with GSEA, p-values were
1157 generated via permutation testing as a fraction of ES from 10,000 permutated gene lists that are
1158 larger than ES observed. Red dotted lines denote the statistical significance threshold (p_BH ≥
1159 0.05). **d** Running-sum profiles for enrichment scores and annotated gene rankings of selected
1160 GO terms between co-expression neighborhoods of *WRKY25* in TEA-GCN and ATTED-II GCN.
1161 In panels d and e, the colors of the dots, running sum profiles, and annotated gene rankings
1162 correspond to the different selected GO terms. Source data for panels a, b, c, and d are provided
1163 in the Source Data file.

1164

1165 **Figure 5. Scalability of TEA-GCN to smaller datasets, its robustness to batch correction,**
1166 **and its GRN inference performance compared to dedicated GRN methods.**

1167 **a** Distribution of AUROC-HM performance scores and AUPRC-HM performance score
1168 measurements between ATTED-II GCNs and *A. thaliana* TEA-GCNs constructed from the
1169 complete transcriptomic dataset (n = 71,720) and randomly downsampled datasets (n = 5,000
1170 and n = 500). In addition, TEA-GCNs were constructed from the subset of samples overlapping
1171 with ATTED-II (14,741, dashed blue violin plot) and from the same subset after ComBat batch
1172 correction (ComBat(14,741*), solid blue violin plot). **b** AUROC-GRN and AUPRC-GRN calculation
1173 relies on GRN positive edges and GRN negative edges. GRN positive edges connect transcription
1174 factors to the targets they activate. GRN positive edges are directional, differing from TF positive
1175 edges; thus, transcription factors targeting each other are connected by two separate positive
1176 edges. GRN negative edges are generated by randomizing genes connected to transcription
1177 factors in GRN positive edges. **b** Distribution of AUROC-GRN and AUPRC-GRN score
1178 measurements of *A. thaliana* TEA-GCNs and GENIE3-GRNs constructed from the complete and
1179 downsampled 14,741 and ComBat(14,741)** transcriptomic datasets, and GRNs inferred by
1180 GENIE3, and GRNBoost2. Non-ensemble PCC GCN is included as a comparison. Receiver
1181 operating characteristic curves (ROCs) and Precision-recall curves (PRCs) correspond to median
1182 performance scores for each method. In panels a and c, each measurement refers to a score
1183 calculated from a set of randomly generated negative edges. In panels a and c, student's t-test
1184 (2-sided) P-values comparing performance measurements of ≤ 0.05, ≤ 0.01, and ≤ 0.001 were
1185 denoted by *, **, and *** respectively, while the adjacent numbers indicate the difference between
1186 medians was shown together with 95% confidence intervals (Difference [Lower interval, Upper
1187 interval]). Exact p-values can be found in source data. Source data for panels a and c are
1188 provided in the Source Data file.

1189

1190 **Figure 6. Annotating co-expression edges of *A. thaliana* TEA-GCN with experimental**
1191 **contexts.**

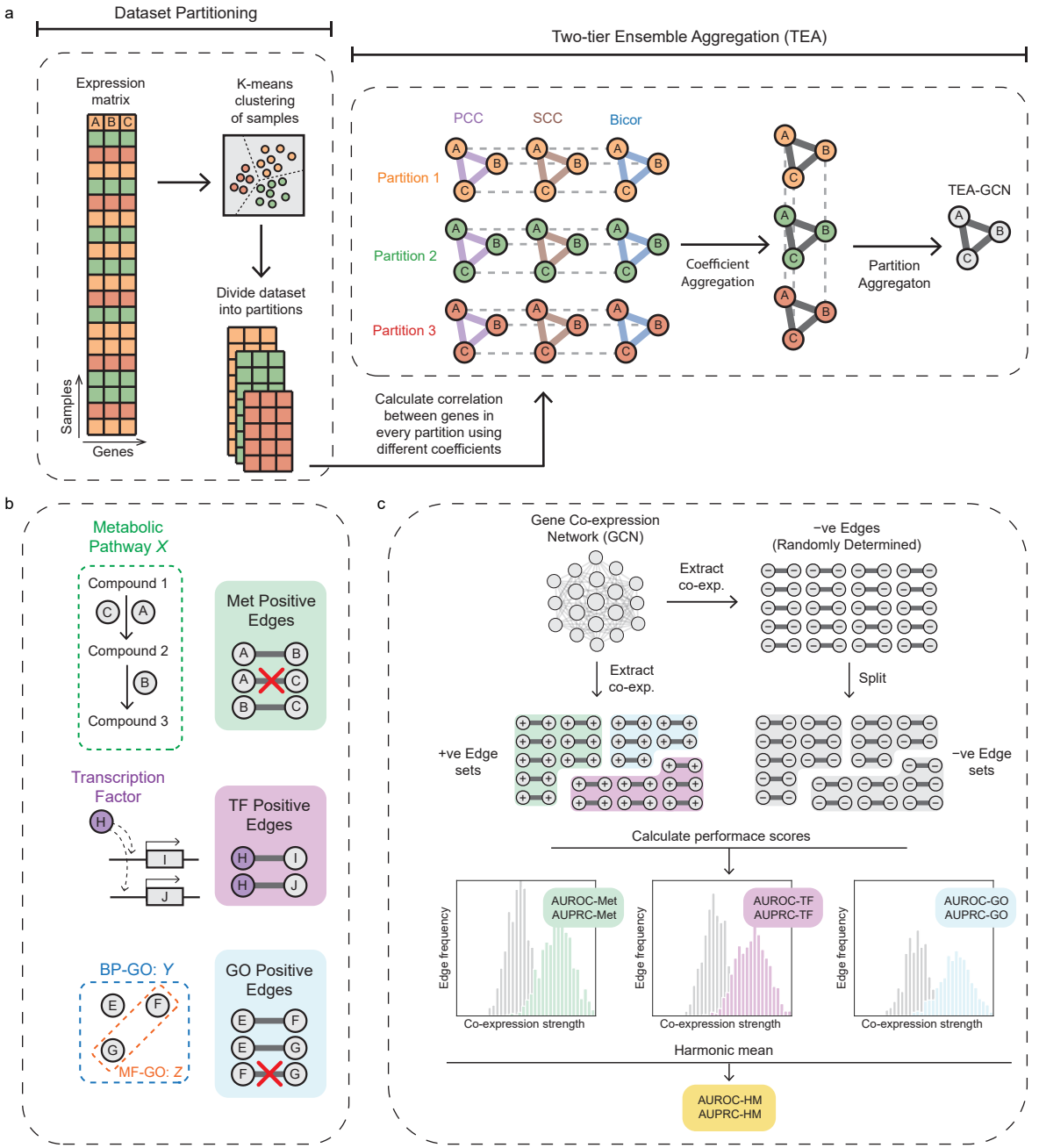
1192 **a** The metadata of RNA-seq samples is broken down into lemmas. Lemmas of RNA-seq samples
1193 belonging to the same dataset partition are then combined. For each partition, occurrences of
1194 different lemmas are counted, and enriched lemmas were determined by comparing the observed
1195 lemma occurrences to permuted occurrences when the partition assignments of RNA-seq
1196 samples are randomly shuffled. Dataset partitions are annotated by overrepresented lemmas

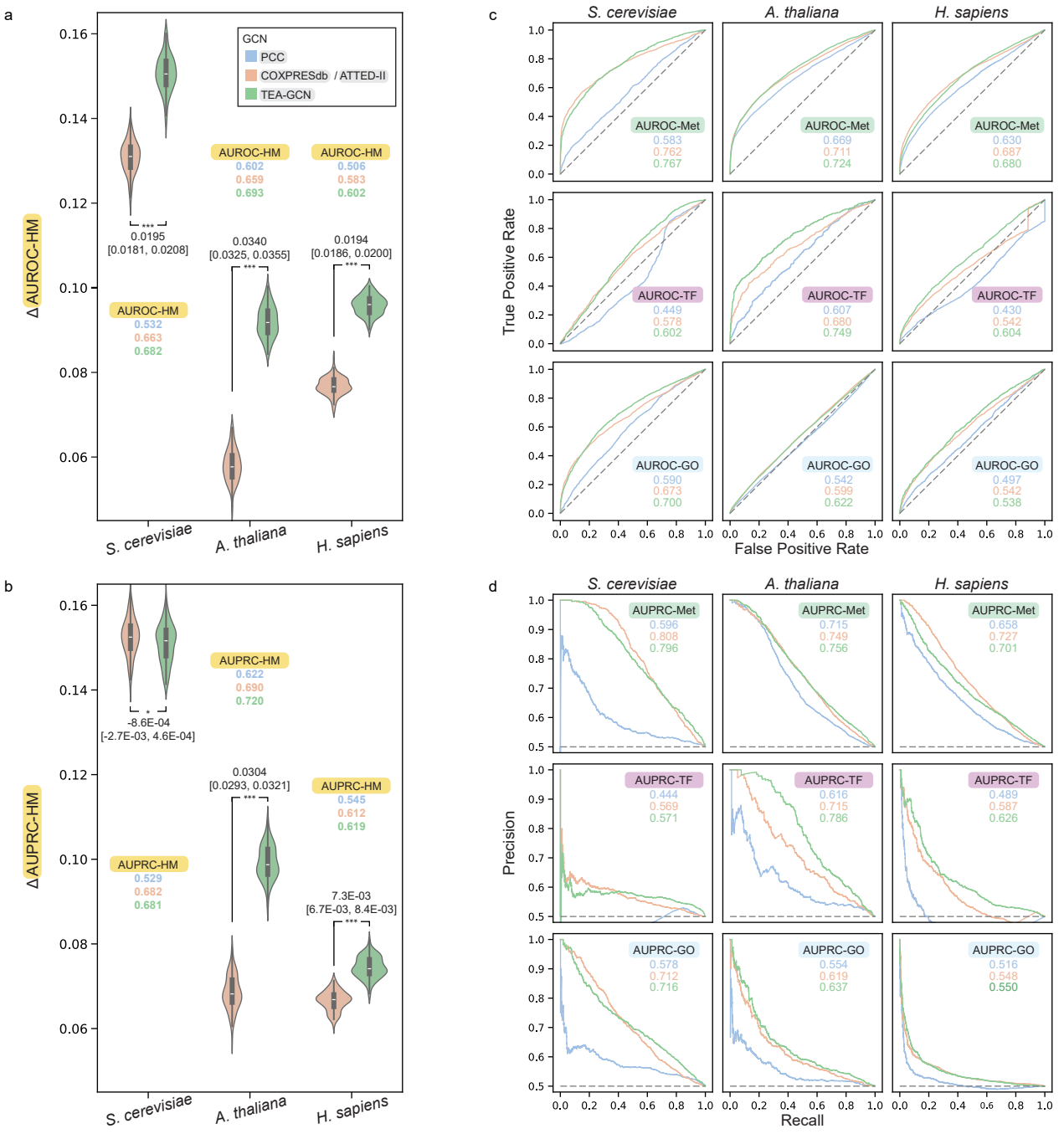
denoting the experimental conditions or tissues from which the samples in the dataset partitions were predominantly generated. **b** Co-expression edges of TEA-GCN can be annotated with lemmas enriched with an approach similar to Gene-set Enrichment Analysis (GSEA). The ensemble co-expression strength 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. Source data for panel c are provided in the Source Data file.

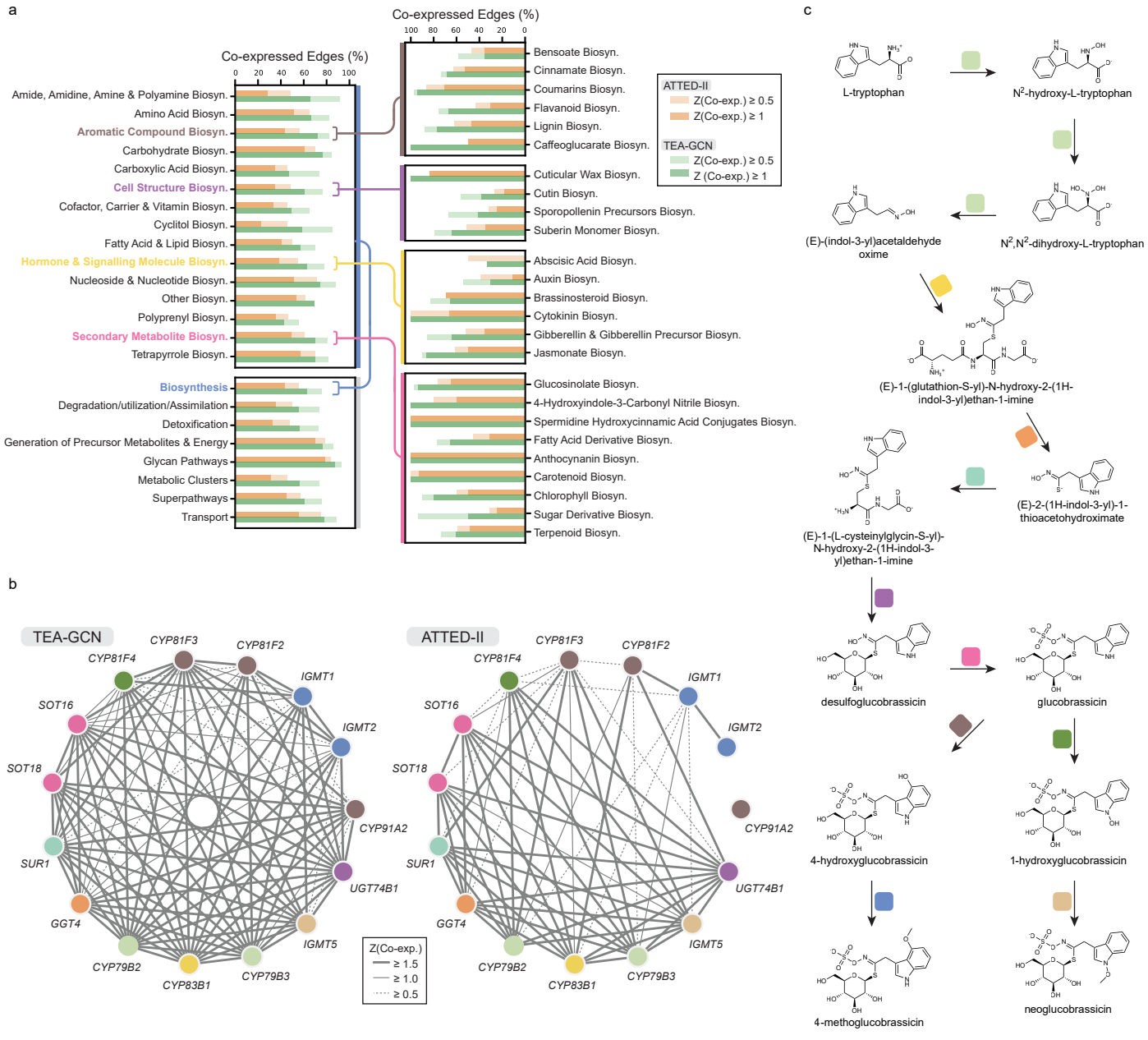
Figure 7. TEA-GCNs show higher cross-species co-expression conservation than non-ensemble PCC GCNs.

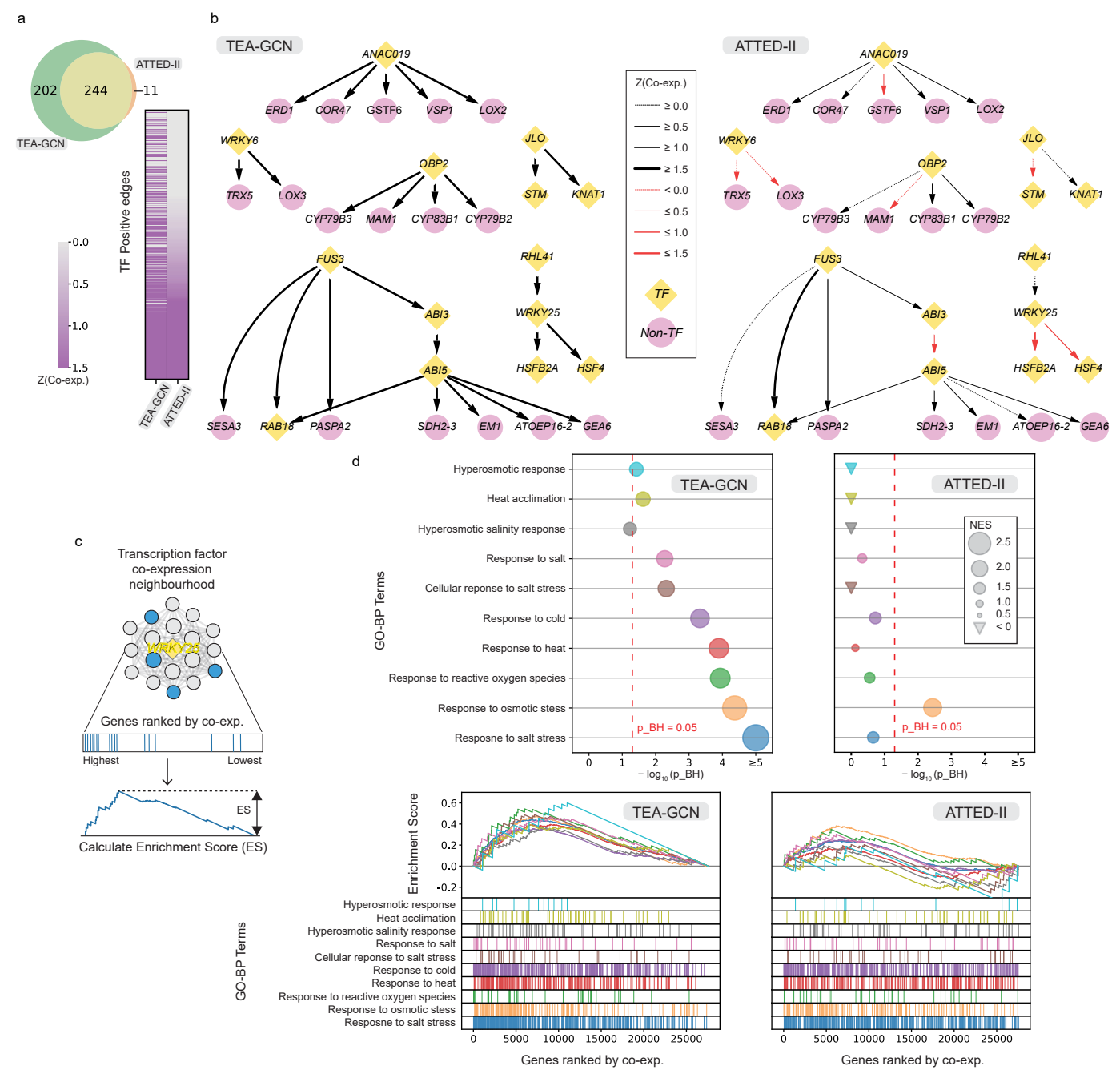
a Images of the ten angiosperm species analyzed. All images are from the public domain (Creative Commons licenses and attributions in Supplementary Data 10). **b** Pairwise conservation of TEA-GCN (left) and PCC GCN (right) ortholog sub-networks across all species. For each method, the upper triangle shows the Spearman's rank correlation of edge weights between ortholog GCNs (co-expression correlation), and the lower triangle shows the Jaccard index of edges above the co-expression threshold (co-expression overlap). Dendrograms summarize the resulting similarity structure between species. **c** Schematic overview of the single-ortholog co-expression conservation analysis. Left: for each species, single-orthologs were inferred using OrthoFinder2 orthology links (Figure S8), retaining only 1-to-1 orthologs and edges connecting them (orthologous edges). Middle: orthologous edges of every species were overlaid to count in how many species each edge is present in and co-expressed in, yielding an edge-level co-expression conservation score. Right: these scores are used to build a single-ortholog co-expression conservation network, in which node-level conservation for each single-ortholog is defined as its weighted degree. **d** Example of a conserved functional module: chloroplast fission (GO:0010020). Nodes correspond to single-ortholog genes, with node colour indicating single-ortholog co-expression conservation, and edges coloured by edge-level co-expression conservation (colour scales on the right). The TEA-GCN-based conservation network (left) forms

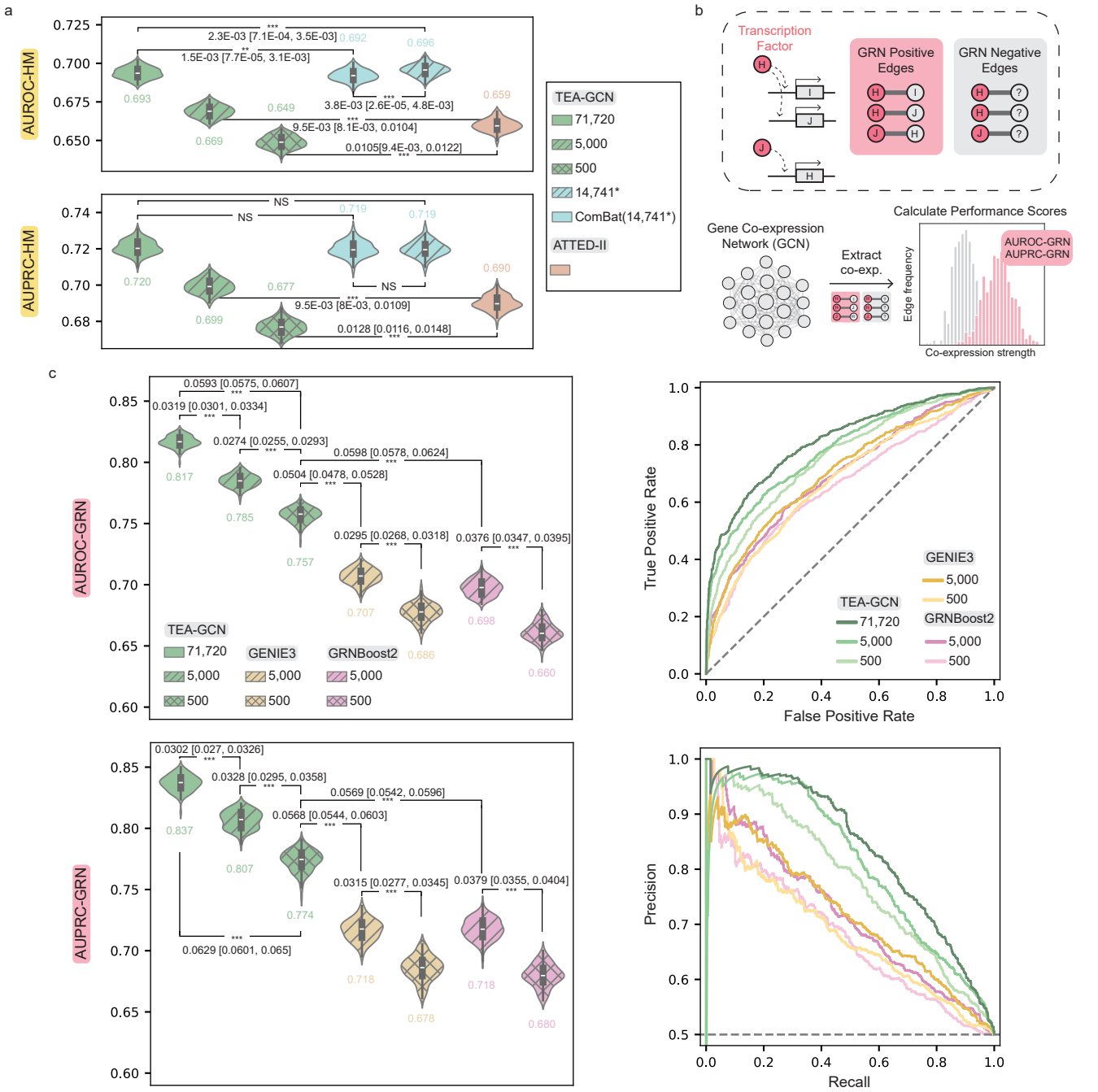
1229 a dense, highly conserved module, whereas the PCC-based conservation network (right) is
1230 markedly weaker and sparser. Source data for panels b and d are provided in the Source Data
1231 file.

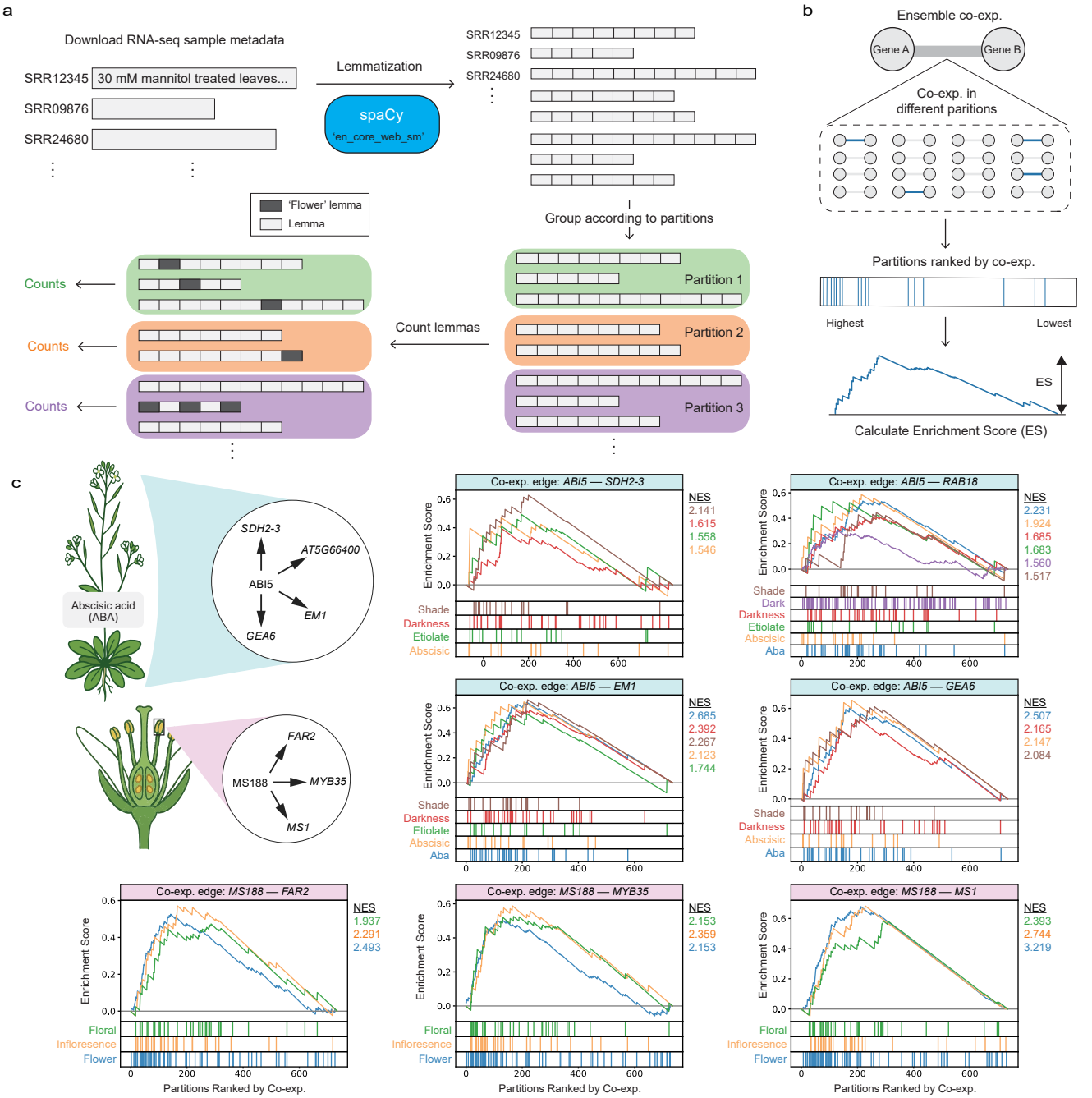

a
Dataset Partitioning
Two-tier Ensemble Aggregation (TEA)
Expression matrix
A B C
Samples
Genes
K-means clustering of samples
Divide dataset into partitions
PCC
SCC
Bicor
Partition 1
Partition 2
Partition 3
Coefficient Aggregation
Partition Aggregaton
TEA-GCN
Calculate correlation between genes in every 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)
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

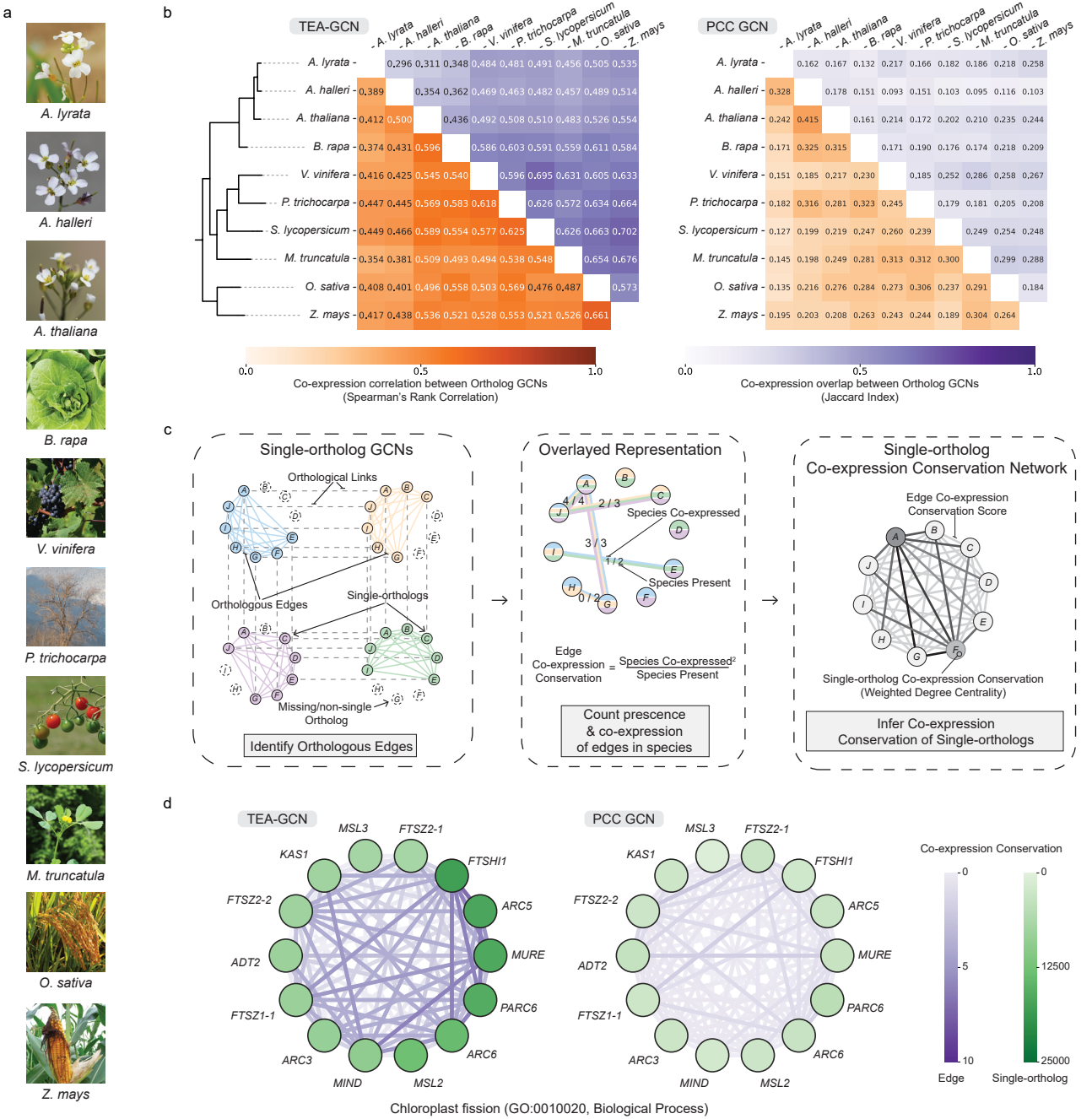

a
GCN
PCC
COXPRESdb / ATTED-II
TEA-GCN
Δ AUROC-HM
AUROC-HM
0.532
0.663
0.682
AUROC-HM
0.602
0.659
0.693
AUROC-HM
0.506
0.583
0.602
0.0195
[0.0181, 0.0208]
0.0340
[0.0325, 0.0355]
0.0194
[0.0186, 0.0200]
S. cerevisiae
A. thaliana
H. sapiens
b
Δ AUPRC-HM
AUPRC-HM
0.529
0.682
0.681
AUPRC-HM
0.622
0.690
0.720
AUPRC-HM
0.545
0.612
0.619
-8.6E-04
[-2.7E-03, 4.6E-04]
0.0304
[0.0293, 0.0321]
7.3E-03
[6.7E-03, 8.4E-03]
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
0.724
AUROC-Met
0.630
0.687
0.680
AUROC-TF
0.449
0.578
0.602
AUROC-TF
0.607
0.680
0.749
AUROC-TF
0.430
0.542
0.604
AUROC-GO
0.590
0.673
0.700
AUROC-GO
0.542
0.599
0.622
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
0.756
AUPRC-Met
0.658
0.727
0.701
AUPRC-TF
0.444
0.569
0.571
AUPRC-TF
0.616
0.715
0.786
AUPRC-TF
0.489
0.587
0.626
AUPRC-GO
0.578
0.712
0.716
AUPRC-GO
0.554
0.619
0.637
AUPRC-GO
0.516
0.548
0.550


a
Co-expressed Edges (%)
0 20 40 60 80 100
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
Bensoate 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.
Anthocynanin 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
b
TEA-GCN
CYP81F3
CYP81F2
CYP81F4
IGMT1
SOT16
IGMT2
SOT18
CYP91A2
SUR1
UGT74B1
GGT4
IGMT5
CYP79B2
CYP83B1
CYP79B3
ATTED-II
CYP81F3
CYP81F2
CYP81F4
IGMT1
SOT16
IGMT2
SOT18
CYP91A2
SUR1
UGT74B1
GGT4
IGMT5
CYP79B2
CYP83B1
CYP79B3
Z(Co-exp.)
≥ 1.5
≥ 1.0
≥ 0.5
c
L-tryptophan
N2-hydroxy-L-tryptophan
N2,N2-dihydroxy-L-tryptophan
(E)-(indol-3-yl)acetaldehyde oxime
(E)-1-(glutathion-S-yl)-N-hydroxy-2-(1H-indol-3-yl)ethan-1-imine
(E)-2-(1H-indol-3-yl)-1-thioacetohydroximate
(E)-1-(L-cysteinylglycin-S-yl)-N-hydroxy-2-(1H-indol-3-yl)ethan-1-imine
desulfoglucobrassicin
glucobrassicin
4-hydroxyglucobrassicin
1-hydroxyglucobrassicin
4-methoglucobrassicin
neoglucobrassicin


a
ATTED-II
202
244
11
TEA-GCN
TF Positive edges
0.0
0.5
1.0
1.5
Z(Co-exp.)
TEA-GCN
ATTED-II
b
TEA-GCN
ATTED-II
ANAC019
ERD1
COR47
GSTF6
VSP1
LOX2
WRKY6
TRX5
LOX3
JLO
STM
KNAT1
OBP2
CYP79B3
MAM1
CYP83B1
CYP79B2
FUS3
RHL41
ABI3
WRKY25
ABI5
HSFB2A
HSF4
SESA3
RAB18
PASPA2
SDH2-3
EM1
ATOEP16-2
GEA6
Z(Co-exp.)
≥ 0.0
≥ 0.5
≥ 1.0
≥ 1.5
< 0.0
≤ 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)
d
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
Resposne to salt stress
p_BH = 0.05
− log10 (p_BH)
NES
2.5
2.0
1.5
1.0
0.5
< 0
Enrichment Score
Genes ranked by co-exp.


a
AUROC-HM
0.725
0.700
0.675
0.650
2.3E-03 [7.1E-04, 3.5E-03]
1.5E-03 [7.7E-05, 3.1E-03]
3.8E-03 [2.6E-05, 4.8E-03]
9.5E-03 [8.1E-03, 0.0104]
0.0105[9.4E-03, 0.0122]
0.693
0.669
0.649
0.692
0.696
0.659
AUPRC-HM
0.74
0.72
0.70
0.68
NS
NS
NS
9.5E-03 [8E-03, 0.0109]
0.0128 [0.0116, 0.0148]
0.720
0.699
0.677
0.719
0.719
0.690
TEA-GCN
71,720
5,000
500
14,741*
ComBat(14,741*)
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
0.0593 [0.0575, 0.0607]
0.0319 [0.0301, 0.0334]
0.0274 [0.0255, 0.0293]
0.0598 [0.0578, 0.0624]
0.0504 [0.0478, 0.0528]
0.0295 [0.0268, 0.0318]
0.0376 [0.0347, 0.0395]
0.817
0.785
0.757
0.707
0.686
0.698
0.660
TEA-GCN
71,720
5,000
500
GENIE3
5,000
500
GRNBoost2
5,000
500
AUPRC-GRN
0.85
0.80
0.75
0.70
0.65
0.60
0.0302 [0.027, 0.0326]
0.0328 [0.0295, 0.0358]
0.0569 [0.0542, 0.0596]
0.0568 [0.0544, 0.0603]
0.0315 [0.0277, 0.0345]
0.0379 [0.0355, 0.0404]
0.0629 [0.0601, 0.065]
0.837
0.807
0.774
0.718
0.678
0.718
0.680
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
GENIE3
5,000
500
TEA-GCN
71,720
5,000
500
GRNBoost2
5,000
500
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
Partition 1
Partition 2
Partition 3
Count lemmas
Counts
Counts
Counts
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
FAR2
MS188
MYB35
MS1
Co-exp. edge: ABI5 — SDH2-3
NES
2.141
1.615
1.558
1.546
Shade
Darkness
Etiolate
Abscisic
Co-exp. edge: ABI5 — RAB18
NES
2.231
1.924
1.685
1.683
1.560
1.517
Shade
Dark
Darkness
Etiolate
Abscisic
Aba
Co-exp. edge: ABI5 — EM1
NES
2.685
2.392
2.267
2.123
1.744
Shade
Darkness
Etiolate
Abscisic
Aba
Co-exp. edge: ABI5 — GEA6
NES
2.507
2.165
2.147
2.084
Shade
Darkness
Abscisic
Aba
Co-exp. edge: MS188 — FAR2
NES
1.937
2.291
2.493
Floral
Infloresence
Flower
Co-exp. edge: MS188 — MYB35
NES
2.153
2.359
2.153
Floral
Infloresence
Flower
Co-exp. edge: MS188 — MS1
NES
2.393
2.744
3.219
Floral
Infloresence
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
TEA-GCN
PCC GCN
Co-expression correlation between Ortholog GCNs
(Spearman's Rank Correlation)
Co-expression overlap between Ortholog GCNs
(Jaccard Index)
c
Single-ortholog GCNs
Orthological Links
Orthologous Edges
Single-orthologs
Missing/non-single Ortholog
Identify Orthologous Edges
Overlayed Representation
Species Co-expressed
Species Present
Edge Co-expression Conservation = Species Co-expressed² / Species Present
Count prescence & co-expression of edges in species
Single-ortholog Co-expression Conservation Network
Edge Co-expression Conservation Score
Single-ortholog Co-expression Conservation (Weighted Degree Centrality)
Infer Co-expression Conservation of Single-orthologs
d
TEA-GCN
PCC GCN
MSL3
FTSZ2-1
KAS1
FTSHI1
FTSZ2-2
ARC5
ADT2
MURE
FTSZ1-1
PARC6
ARC3
ARC6
MIND
MSL2
Co-expression Conservation
Edge
Single-ortholog
Chloroplast fission (GO:0010020, Biological Process)