Supplementary Information for:

Biological pathway expression complementation contributes to biomass heterosis in *Arabidopsis*

Wenwen Liu[a], Guangming He[a,1], Xing Wang Deng[a,b,1]

[a]School of Advanced Agricultural Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

[b]Peking University-Southern University of Science and Technology Institute of Plant and Food Science, Department of Biology, Southern University of Science and Technology, Shenzhen 518055, China.

[1]Corresponding authors: Guangming He (heguangming@pku.edu.cn); Xing Wang Deng (deng@pku.edu.cn).

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Datasets S1 to S11

Materials and Methods

Plant materials and sampling

We obtained *A. thaliana* accessions Col-0 and Per-1 from the *Arabidopsis* Biological Resource Center and used hand pollination with Col-0 as the maternal line to generate Col-0 × Per-1 F_1 hybrid seeds. The plants used for phenotyping and transcriptomic analyses were grown on Murashige & Skoog plates containing 1% sucrose under long-day conditions (16-h light at 18 W/m^2 and 22°C and 8-h dark at 18°C) after the seeds had been surface-sterilized and stratified for 7 d. For RNA-seq, we grew the parental lines alongside the hybrids in the same plate so that each plate was a biological replicate, and all plates in the light incubator were rotated twice a day. In the mornings, plant shoots were harvested at 3–8 DAS, and the first or second true leaf from each plant was harvested at 7–21 DAS. We set up three biological replicates for each time point and each genotype, with each replicate comprised of samples collected from at least 10 plants. For phenotyping, plants were grown under the same conditions as for RNA-seq and the tissues were sampled in the mornings at 3–21 DAS with at least five biological replicates.

Phenotyping

The sampled tissues were bleached with ethanol and then placed in chloral hydrate solution for clearing (1). After obtaining images of the cotyledons and true leaves using microscopes, we measured the cotyledon and leaf sizes using ImageJ (https://imagej.nih.gov/ij/).

RNA sequencing and data processing

We used an RNeasy Plant Mini Kit with on-column DNase treatment (Qiagen) to extract total RNA from each of the 189 samples. Next, we constructed mRNA sequencing libraries and sequenced the samples on a HiSeq X Ten platform (Illumina) to generate 150-nucleotide paired-end reads. Each sample yielded approximately 29,000,000 to 59,000,000 raw reads (Dataset S2). Quality control was conducted using fastp version 0.20.0 (2) with the parameter "length_required" set to 150 to generate approximately 7,600,000,000 high-quality reads for all RNA-seq samples, and yielding approximately 28,000,000 to 57,000,000 for each sample.

Using Salmon version 1.0.0 (3), we quantified transcript expression against a high-quality modified version of the *Arabidopsis Thaliana* Reference Transcript Dataset 2 (AtRTDv2_QUASI_19April2016), which is specifically designed to be used with Salmon to accurately quantify alternatively spliced isoforms; it contains 81,620 non-redundant transcripts from 33,681 genes (4). We built a mapping-based index in a default type ("puff") using an auxiliary k-mer hash over k-mers with a length of 31, and then we quantified reads directly against this index using the mapping-based mode in Salmon while correcting for sequence-specific biases with the option --seqBias. The number of bootstrap samples to compute was set to 30, and the "--validateMappings" flag was passed to enable selective alignment, which can improve the accuracy of both mapping and quantification estimates. All other options were set to default (3). The number of mapped equivalence reads and the mapping rate are listed in Dataset S2.

The transcript-level quantifications were merged to the gene level, and then length-scaled transcripts per million (TPM) and estimated read counts were calculated for genes using R package tximport version 1.12.3 with the option lengthScaledTPM to correct for possible gene length variations across samples (5). We then removed genes expressed at very low levels, defining a low-expressed gene as having no transcripts with ≥ 1 counts per million in 3 or more of the 189 samples. A total of 20,427 expressed genes were used for downstream analysis.

PCA and sample correlation analysis

Gene expression values (TPM) for the top 1,000 genes with the highest standard deviations across the 189 RNA-seq samples were used for PCA using the prcomp function in R, with default settings. We also calculated Pearson correlation coefficients between pairwise samples with the TPM of all 20,427 expressed genes.

Differential gene expression analysis

For each time point within the early shoot and true leaf transcriptomes, we compared the differential gene expression between the three genotypes (Col-0, Per-1, and hybrid) against each other. To calculate differential expression for pairwise tests, we modeled the read counts of the expressed genes using the linear fit function voom in R package limma version 3.40.6 to correct for library size differences. The limma empirical Bayes function was used to identify significant differential expression of genes (6, 7), and these genes were filtered to retain only those with 1) an average of ≥ 1 TPM of biological replicates in at least one of the pairwise groups and 2) an adjusted *p*-value < 0.05. In general, we classified the genes into 14 categories (Dataset S11), including patterns of above parent expression, above high-parent expression, high-parent expression, below parent expression, below low-parent expression, low-parent expression, and additive expression in the hybrid compared to its parents.

Weighted gene coexpression network analysis (WGCNA)

For transcriptome atlases of 3–8 DAS shoots, WGCNA was performed for the individual gene expression dataset of each genotype using the functions in R package WGCNA version 1.68 (8). Each individual dataset contained the $\log_2$(TPM + 1) of genes that had an average of ≥ 1 TPMs of biological replicates in at least one time point during the 3–8 DAS period for the corresponding genotype and were within the top 75% of the above genes that had the highest median absolute deviation of $\log_2$(TPM + 1) across different time points. Each of these datasets underwent network analysis with a soft threshold (power/β) that was determined to produce a scale-free network with optimal scale-free topology model fit and mean connectivity (*SI Appendix*, Fig. S3 and Table S2). Next, we used the WGCNA blockwiseModules function to construct a signed network. Briefly, gene coexpression relationships were calculated as bi-weight mid-correlation coefficients raised to the soft threshold, transforming the gene expression correlation adjacency matrix to a TOM, which was then converted to a dissimilarity matrix that was used to generate a hierarchical cluster tree. To identify the coexpressed gene modules, we used the dynamic tree cut method with the following parameters: deepSplit level 2, detectCutHeight of 0.995, minModuleSize of 100, and tree mergeCutHeight of 0.25.

We performed module preservation analysis among the individually constructed gene coexpression networks of the three genotypes using the WGCNA modulePreservation function with nPermutations of 50 and networkType “signed” (9). The permutation test defined the preservation degree of a module in 2 networks by providing a $Z_{summary}$ value that summarized density- and connectivity-based preservation statistics, where $Z_{summary} < 2$ represented no preservation, $2 < Z_{summary} < 10$ represented weak to moderate preservation, and $Z_{summary} > 10$ represented strong preservation.

Using the function blockwiseConsensusModules in WGCNA, we detected the conserved gene coexpression network (i.e., the consensus modules) underlying early shoot development among the three genotypes (10, 11). Briefly, a consensus adjacency matrix was created using the scaled adjacency matrices from each individual dataset with consensusQuantile set to 0, and then a consensus TOM was generated from the consensus adjacency matrix. Consensus modules were calculated using hierarchical clustering and dynamic tree cutting with the following parameters: deepSplit level 2, detectCutHeight of 0.99, minModuleSize of 50, and mergeCutHeight of 0.25. Modules presenting both high correlation and similar expression profiles were merged using the WGCNA mergeCloseModules function with cutHeight set to 0.25 and consensusQuantile set to 0.25. To examine the gene expression patterns of the consensus modules across samples, each module was represented by an module eigengene (ME), which was calculated as the first principle component of the expression profiles of each module (8). The connectivity of each gene to its corresponding module was calculated using a module membership (k_{ME}) value that was defined as the bi-weight mid-correlation between the gene expression and the corresponding ME (8). Using each gene's intramodular k_{ME}, we found hub genes in the conserved network of each genotype (8, 12). The relationships between consensus coexpression modules in each genotype were studied by examining the eigengene network, which was built using the bi-weight mid-correlation between module eigengenes (10).

Similarly, we generated a gene expression dataset of 7–21 DAS leaf transcriptomes for WGCNA of each genotype. The construction and analysis of individual and consensus gene coexpression networks were performed as described above (*SI Appendix*, Fig. S5 and Table S2).

Gene annotation and ontology enrichment analysis

Gene descriptions and GO terms for *A. thaliana* were assigned according to The *Arabidopsis* Information Resource (https://www.arabidopsis.org/) and R package org.At.tair.db. The GO enrichment analysis was performed with GO Biological Process Complete using the R package clusterProfiler version 3.10 (13) and applying a hypergeometric test with FDR correction (adjusted $P < 0.05$).

Data availability

All code used to perform RNA-seq analysis and WGCNA is publicly available on github at https://github.com/WenwenLiu54. All original transcriptome sequences and gene expression data have been deposited in the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo) under accession number GSE157957. Additional data, such as raw image files, that support this study are available from the corresponding authors upon request.

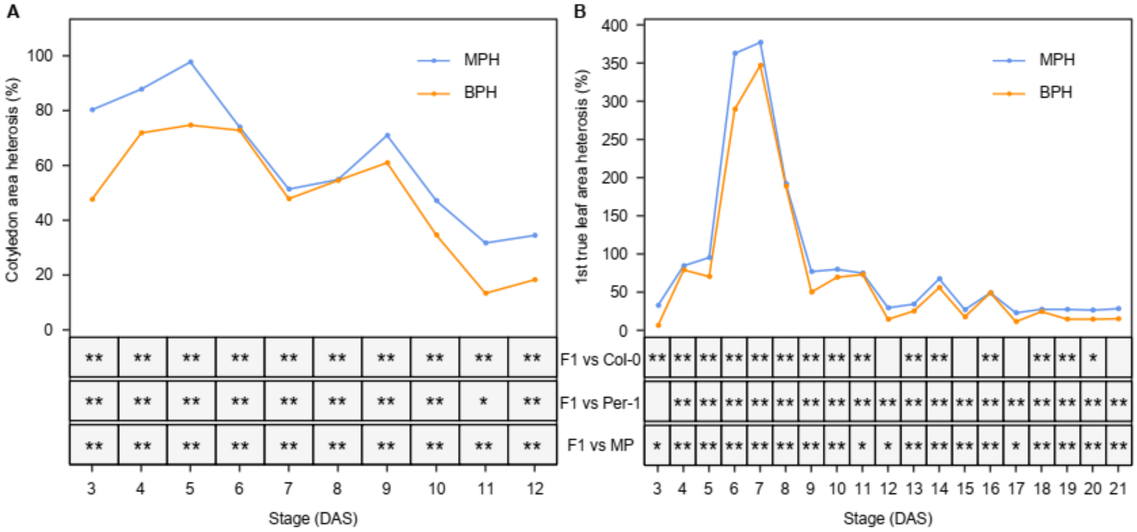


Fig. S1. Cotyledon and true leaf growth heterosis in Col-0 × Per-1 during seedling development. Mid-parent heterosis (MPH) and best-parent heterosis (BPH) levels of the cotyledon area at 3–12 days after sowing (DAS) (*A*) and the first true leaf area at 3–21 DAS (*B*) in *Arabidopsis* Col-0 × Per-1. Asterisks (*) indicate significant differences between the F_1 hybrid and Col-0, Per-1, and MP (mid-parent, the average level of both parents). ***P* < 0.01, **P* < 0.05; Student's *t*-test, *n* ≥ 5.

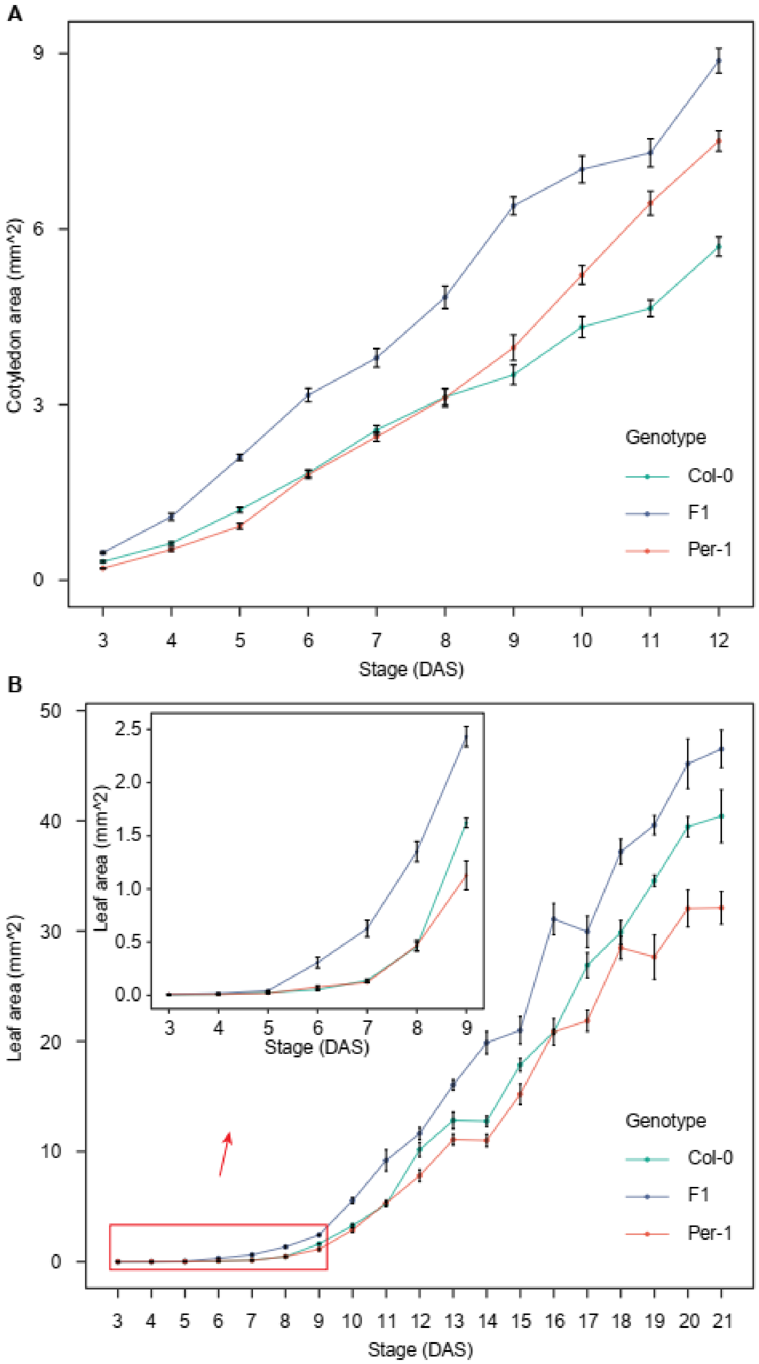


Fig. S2. Dynamic growth of cotyledon at 3–12 days after sowing (DAS) (*A*) and the first true leaf at 3–21 DAS (*B*) in Col-0, Per-1, and Col-0 × Per-1. Data are presented as mean area ± standard error in each genotype at each stage ($n \geq 5$).

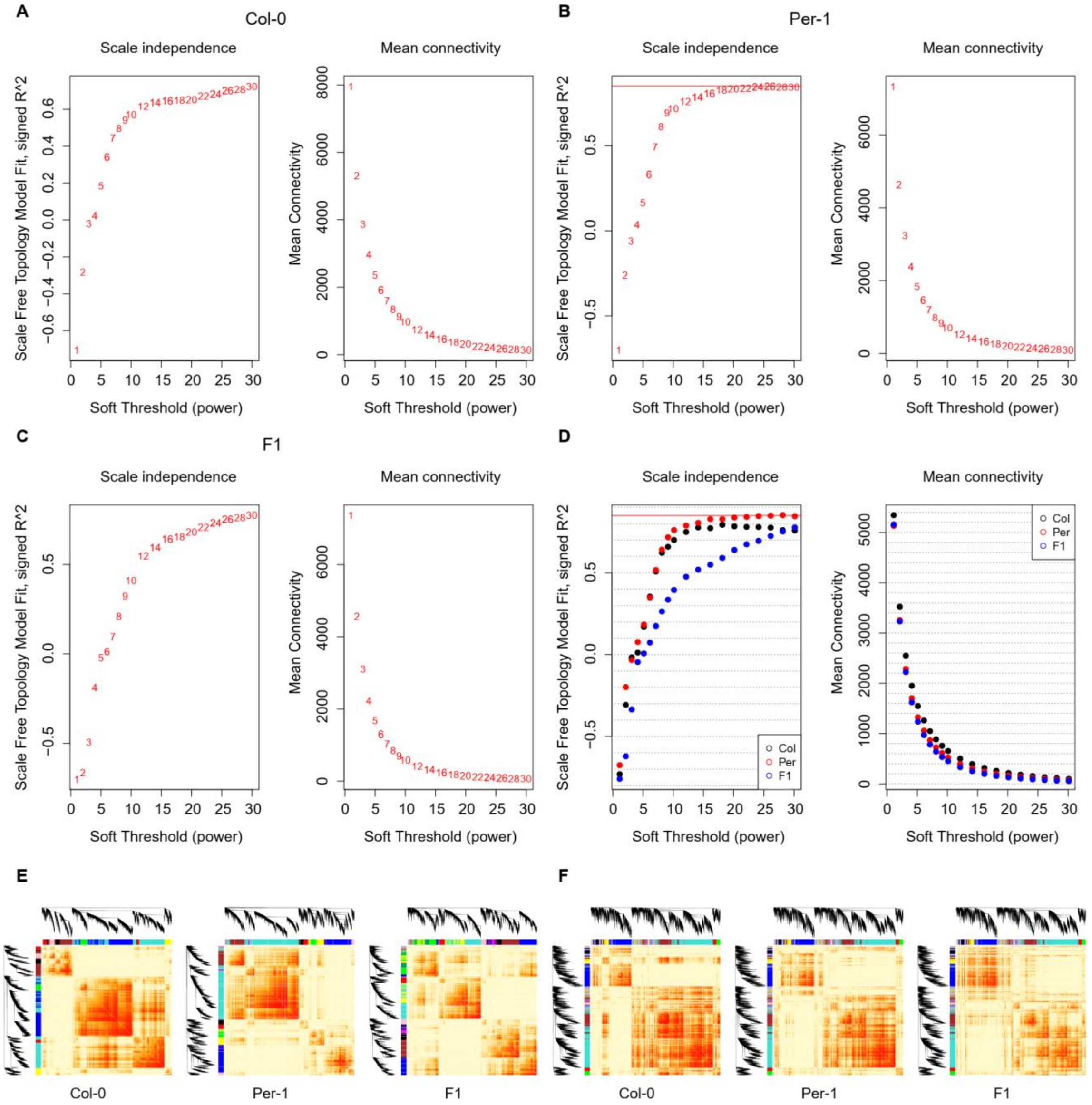


Fig. S3. Determination of soft thresholds (power/β) that provided optimal scale-free topology indices of gene coexpression networks underlying early shoot development (3–8 DAS) and the corresponding topological overlap matrices (TOMs) for gene expression correlations. (*A*-*C*) Network scale-free topology model fit and mean connectivity under different soft thresholds for individual gene expression datasets of ecotypes Col-0 (*A*), Per-1 (*B*), and their F_1 hybrid (*C*). (*D*) Network scale-free

topology model fit and mean connectivity under different soft thresholds for integrated gene expression datasets in Col-0, Per-1, and the hybrid. (*E* and *F*) Heat maps showing TOMs for gene expression correlations (soft threshold =18) in Col-0, Per-1, and the hybrid used for individual gene coexpression network construction (*E*) and consensus gene coexpression network construction (*F*). See *SI Appendix*, Table S2.

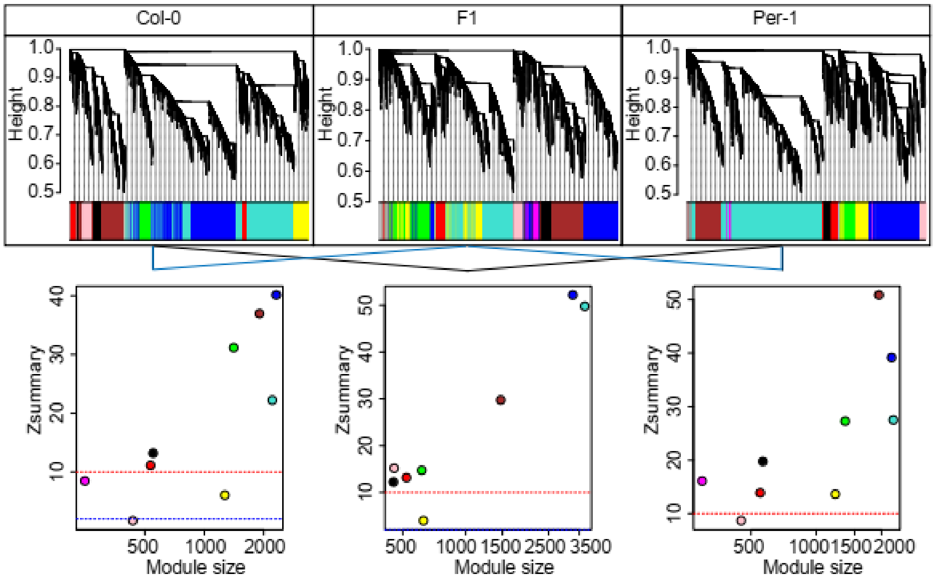


Fig. S4. Preservation of gene coexpression networks between the hybrid and its parents underlying early shoot development (3–8 DAS). Hierarchical cluster dendrograms showing gene coexpression modules identified by weighted gene coexpression network analysis (WGCNA) of *Arabidopsis* ecotypes (Col-0, Per-1) and their F$_1$ hybrid. In the dendrograms, each leaf represents one gene and each module below the dendrograms is labeled with one color. The genes without coexpression with any module are marked in grey. Below the dendrograms are pairwise module preservation analyses of the networks of the three genotypes. Dashed red and blue lines represent the $Z_{summary}$ thresholds for strong ($Z_{summary} > 10$) and weak to moderate ($2 < Z_{summary} < 10$) preservation levels, respectively. Colored dots represent the corresponding modules in the reference network, and the module size is the number of overlapped genes within each reference module.

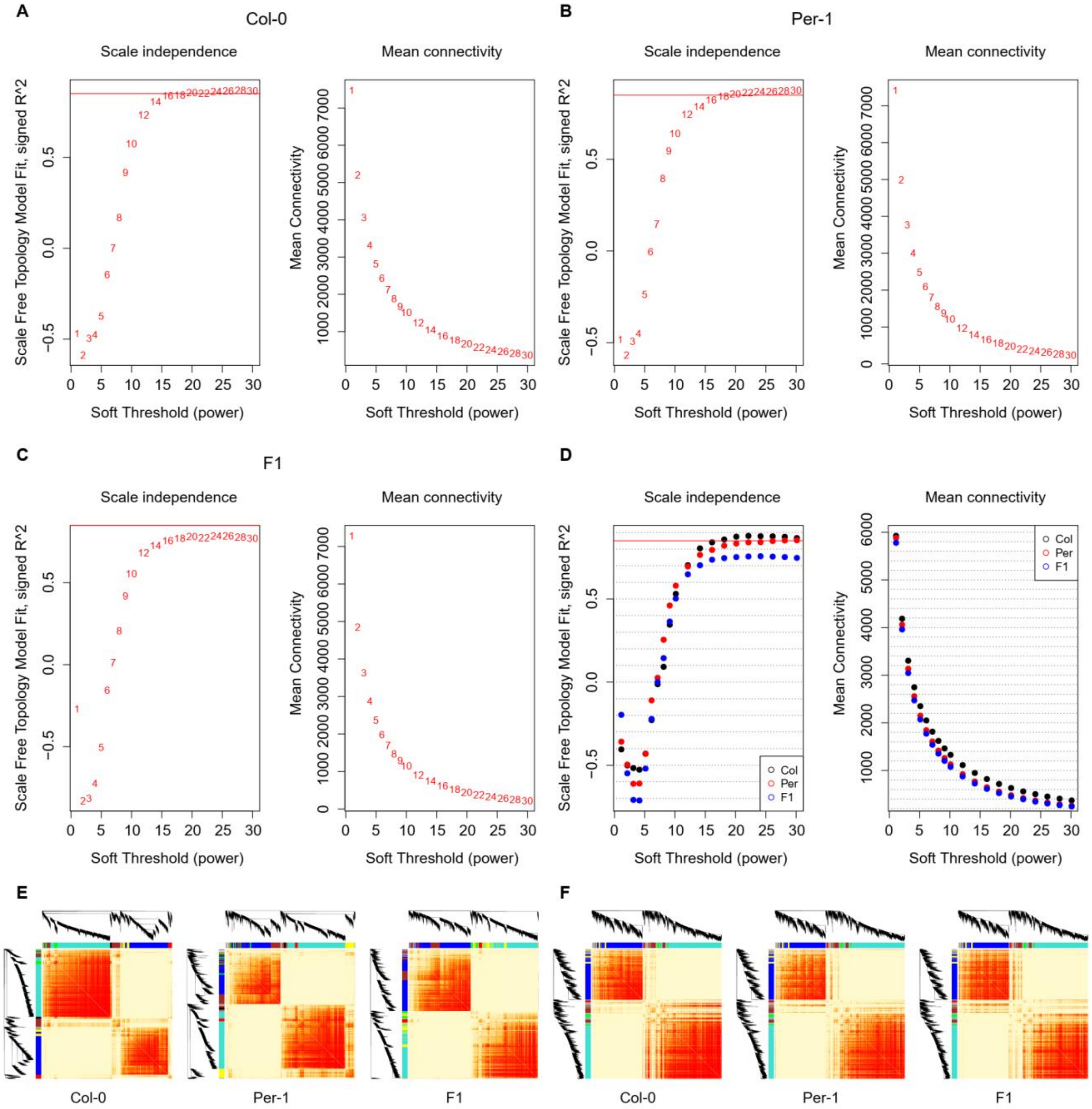


Fig. S5. Determination of soft thresholds (power/β) that provided optimal scale-free topology indices of gene coexpression networks underlying true leaf development (7–21 DAS) and the corresponding topological overlap matrices (TOMs) for gene expression correlations. (*A*-*C*) Network scale-free topology model fit and mean connectivity under different soft thresholds for individual gene expression datasets of ecotypes Col-0 (*A*), Per-1 (*B*), and their F_1 hybrid (*C*). (*D*) Network scale-free

topology model fit and mean connectivity under different soft thresholds for integrated gene expression datasets in Col-0, Per-1, and the hybrid. (*E* and *F*) Heat maps showing TOMs for gene expression correlations (soft threshold = 20) in Col-0, Per-1, and the hybrid used for individual gene coexpression network construction (*E*) and consensus gene coexpression network construction (*F*). See *SI Appendix*, Table S2.

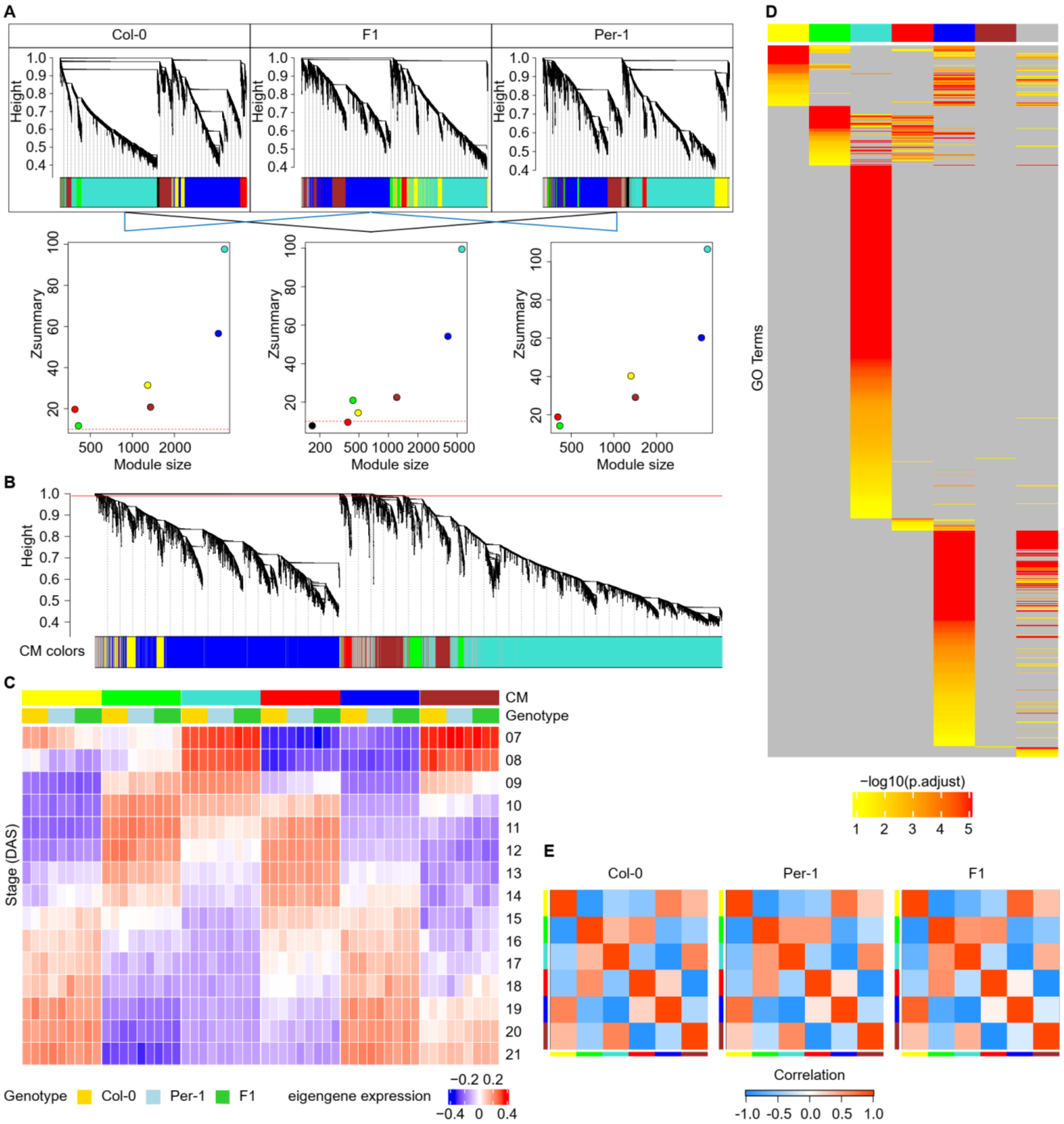
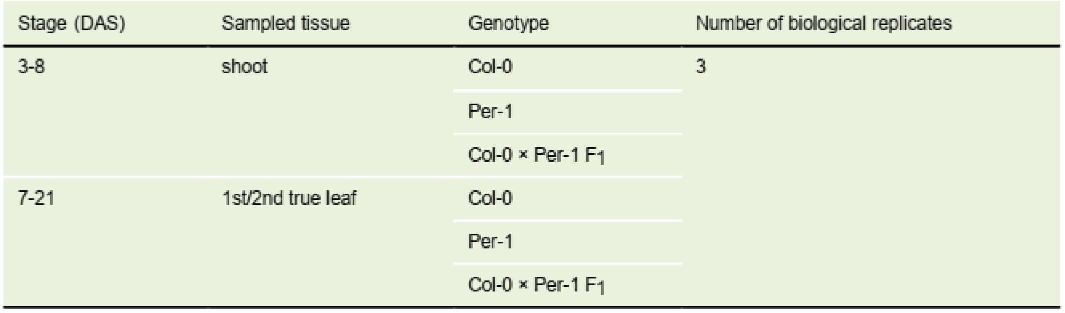


Fig. S6. Preservation and divergence of gene coexpression networks between hybrid and parents underlying true leaf development (7–21 DAS). (*A*) Hierarchical cluster dendrograms showing gene coexpression modules identified by weighted gene coexpression network analysis in ecotypes (Col-0, Per-1) and their F$_1$ hybrid. In the dendrograms, each leaf represents one gene and each module is labeled with one color. The genes without coexpression with any module are marked in grey. Pairwise module

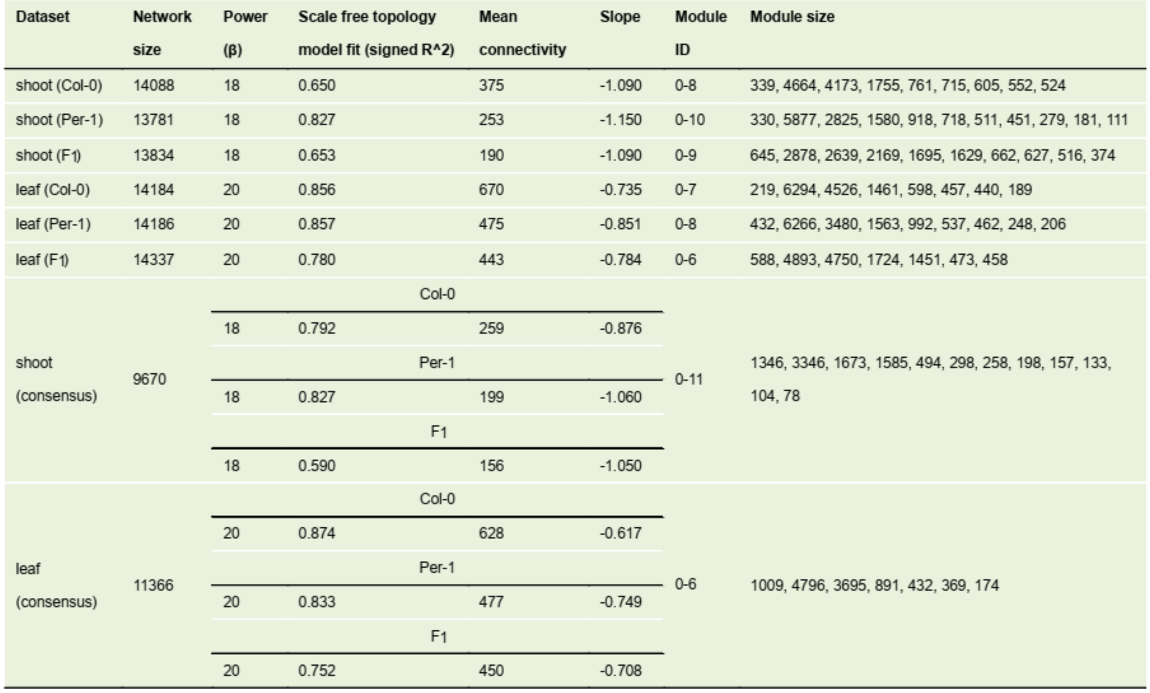
preservation analyses of the networks in the three genotypes appear below the dendrograms. Dashed red and blue lines represent the Z_{summary} thresholds for strong ($Z_{\text{summary}} > 10$) and weak to moderate ($2 < Z_{\text{summary}} < 10$) preservation levels, respectively. Colored dots represent the corresponding modules in the reference network, and the module size is the number of overlapped genes within each reference module. (*B*) Hierarchical cluster dendrogram showing consensus gene coexpression modules. The red line above is the cut height (0.99) for consensus module (CM) identification. CM colors are relabeled independent of the separately identified modules in each genotype, and the genes not coexpressed in all three genotypes are marked in grey. (*C*) Differential eigengene expression patterns of the CMs in the three genotypes across 7–21 DAS. Blue and red represent lesser and greater expression, respectively. (*D*) Heat map summarizing GO enrichment of genes in each CM. The color intensity represents the significance (–$\log_{10}$[adjusted *p*-value]) of GO term enrichment. (*E*) Heat map of eigengene networks showing relationships among CMs in each genotype.

Table S1. Overview of the high temporal resolution samples used for RNA-seq analysis



Stage (DAS)	Sampled tissue	Genotype	Number of biological replicates
3-8	shoot	Col-0	3
		Per-1	
		Col-0 × Per-1 F_1	
7-21	1st/2nd true leaf	Col-0	
		Per-1	
		Col-0 × Per-1 F_1	

Table S2. Parameters and properties of individual and consensus gene coexpression networks



Dataset	Network size	Power (β)	Scale free topology model fit (signed R^2)	Mean connectivity	Slope	Module ID	Module size
shoot (Col-0)	14088	18	0.650	375	-1.090	0-8	339, 4664, 4173, 1755, 761, 715, 605, 552, 524
shoot (Per-1)	13781	18	0.827	253	-1.150	0-10	330, 5877, 2825, 1580, 918, 718, 511, 451, 279, 181, 111
shoot (F_1)	13834	18	0.653	190	-1.090	0-9	645, 2878, 2639, 2169, 1695, 1629, 662, 627, 516, 374
leaf (Col-0)	14184	20	0.856	670	-0.735	0-7	219, 6294, 4526, 1461, 598, 457, 440, 189
leaf (Per-1)	14186	20	0.857	475	-0.851	0-8	432, 6266, 3480, 1563, 992, 537, 462, 248, 206
leaf (F_1)	14337	20	0.780	443	-0.784	0-6	588, 4893, 4750, 1724, 1451, 473, 458
shoot (consensus)	9670	Col-0				0-11	1346, 3346, 1673, 1585, 494, 298, 258, 198, 157, 133, 104, 78
		18	0.792	259	-0.876		
		Per-1					
		18	0.827	199	-1.060		
		F_1					
		18	0.590	156	-1.050		
leaf (consensus)	11366	Col-0				0-6	1009, 4796, 3695, 891, 432, 369, 174
		20	0.874	628	-0.617		
		Per-1					
		20	0.833	477	-0.749		
		F_1					
		20	0.752	450	-0.708		

Dataset S1 (XLS). Dynamic growth and heterosis of the cotyledon and the first true leaf in Col-0, Per-1, and the hybrid.

Dataset S2 (XLS). Overview of the RNA-seq raw data and mapping information.

Dataset S3 (XLS). GO enrichment information for the genes in each CM of the conserved gene coexpression network underlying early shoot development. Significance is presented as the adjusted *p*-value. NS, no significance.

Dataset S4 (XLS). Overlap of the hub genes identified in the core regulatory network underlying early shoot development among Col-0, Per-1, and the hybrid.

Dataset S5 (XLS). GO enrichment information for all hub genes of the core regulatory network underlying early shoot development identified in three genotypes.

Dataset S6 (XLS). Differential expression patterns (adjusted *p*-value < 0.05) and function descriptions of hub genes involved in the mitotic cell cycle and photosynthesis underlying early shoot development.

Dataset S7 (XLS). GO enrichment information for genes in each CM of the conserved gene coexpression network underlying true leaf development. Significance is presented as the adjusted *p*-value. NS, no significance.

Dataset S8 (XLS). Overlap of the hub genes identified in the core regulatory network underlying true leaf development among Col-0, Per-1, and the hybrid.

Dataset S9 (XLS). GO enrichment information for all hub genes of the core regulatory network underlying true leaf development identified in three genotypes.

Dataset S10 (XLS). Differential expression patterns (adjusted *p*-value < 0.05) and function descriptions of hub genes involved in the mitotic cell cycle and photosynthesis underlying true leaf development.

Dataset S11 (XLS). The 14 categories of differential gene expression patterns identified in the hybrid compared to its parents (adjusted *p*-value < 0.05).

SI References

1. B. Rymen, F. Coppens, S. Dhondt, F. Fiorani, G. T. Beemster, Kinematic analysis of cell division and expansion. *Methods Mol Biol* **655**, 203-227 (2010).
2. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
3. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
4. R. Zhang *et al.*, A high quality *Arabidopsis* transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic acids research* **45**, 5061-5073 (2017).
5. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).
6. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15**, R29 (2014).
7. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015).
8. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559 (2008).

9. P. Langfelder, R. Luo, M. C. Oldham, S. Horvath, Is my network module preserved and reproducible? *PLoS Comput Biol* **7**, e1001057 (2011).

10. P. Langfelder, S. Horvath, Eigengene networks for studying the relationships between co-expression modules. *Bmc Syst Biol* **1**, 54 (2007).

11. M. Zinkgraf, L. Liu, A. Groover, V. Filkov, Identifying gene coexpression networks underlying the dynamic regulation of wood-forming tissues in *Populus* under diverse environmental conditions. *The New phytologist* **214**, 1464-1478 (2017).

12. P. Langfelder, P. S. Mischel, S. Horvath, When is hub gene selection better than standard meta-analysis? *PloS one* **8**, e61505 (2013).

13. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).