G3, 2024, jkae059
https://doi.org/10.1093/g3journal/jkae059
Advance Access Publication Date: 16 March 2024
Plant Genetics and Genomics

Harnessing the predicted maize pan-interactome for putative gene function prediction and prioritization of candidate genes for important traits

Elly Poretsky,[1,†] Halise Busra Cagirici,[1,5,†] Carson M. Andorf ,[2,3] Taner Z. Sen [1,4,*]

[1]Crop Improvement and Genetics Research Unit, U.S. Department of Agriculture, Agricultural Research Service, 800 Buchanan St., Albany, CA 94710, USA
[2]Corn Insects and Crop Genetics Research, U.S. Department of Agriculture, Agricultural Research Service, Ames, IA 50011, USA
[3]Department of Computer Science, Iowa State University, Ames, IA 50011, USA
[4]Department of Bioengineering, University of California, 306 Stanley Hall, Berkeley, CA 94720, USA
[5]Present address: Division of Infectious Diseases and Geographic Medicine, Department of Medicine, Stanford University School of Medicine, 300 Pasteur Drive, L-13, Stanford, CA 94305, USA

*Corresponding author: Crop Improvement and Genetics Research Unit, U.S. Department of Agriculture, Agricultural Research Service, 800 Buchanan St., Albany, CA 94710, USA. Email: taner.sen@usda.gov
[†]These authors contributed equally to this work.

The recent assembly and annotation of the 26 maize nested association mapping population founder inbreds have enabled large-scale pan-genomic comparative studies. These studies have expanded our understanding of agronomically important traits by integrating pan-transcriptomic data with trait-specific gene candidates from previous association mapping results. In contrast to the availability of pan-transcriptomic data, obtaining reliable protein–protein interaction (PPI) data has remained a challenge due to its high cost and complexity. We generated predicted PPI networks for each of the 26 genomes using the established STRING database. The individual genome-interactomes were then integrated to generate core- and pan-interactomes. We deployed the PPI clustering algorithm ClusterONE to identify numerous PPI clusters that were functionally annotated using gene ontology (GO) functional enrichment, demonstrating a diverse range of enriched GO terms across different clusters. Additional cluster annotations were generated by integrating gene coexpression data and gene description annotations, providing additional useful information. We show that the functionally annotated PPI clusters establish a useful framework for protein function prediction and prioritization of candidate genes of interest. Our study not only provides a comprehensive resource of predicted PPI networks for 26 maize genomes but also offers annotated interactome clusters for predicting protein functions and prioritizing gene candidates. The source code for the Python implementation of the analysis workflow and a standalone web application for accessing the analysis results are available at https://github.com/eporetsky/PanPPI.

Keywords: Plant Genetics and Genomics; pan-genome; predicted protein–protein interactions (PPIs); interactome; protein function; gene candidate prioritization

Introduction

Maize (*Zea mays*) is one of the most agriculturally and economically important crops in the world (Hufford *et al.* 2021). In an effort to improve yield and reduce loss to stress conditions, association mapping for important agronomic traits has been extensively used to better understand the genetic basis underlying phenotypic differences across the genomic diversity of maize (Wallace *et al.* 2014; Mural *et al.* 2022). One major mapping population, the maize nested association mapping (NAM) population, consists of the products of crosses between 25 diverse founder inbred lines and the B73 reference genome inbred line that represent a large portion of maize genetic diversity (McMullen *et al.* 2009; Hufford *et al.* 2021). Association studies conducted using the NAM mapping population identified a large number of genomic loci and gene candidates associated with a variety of traits, such as plant architecture, height, flowering time, kernel weight, and different metabolite abundances (Buckler *et al.* 2009; Peiffer *et al.* 2014; Wallace *et al.* 2014; Pan *et al.* 2017; Zhang *et al.* 2020). To better understand the genetic and molecular basis of these traits will require improvements in gene function prediction and prioritization of causal candidate genes (Visscher *et al.* 2017). Thus, despite the comprehensive understanding of the genetic architecture and the association between some traits and genomic loci, identification of the causal genes and the underlying biological networks regulating their function remains elusive for many other traits (Broekema *et al.* 2020). Such identification would facilitate both crop improvement and progress in understanding complex biological systems.

High-quality genome assemblies of diverse plant species revealed a more complete picture of the biological regulations and traits of agronomic importance (Kersey 2019; Sun *et al.* 2022; Shi *et al.* 2023). Improvements in the quality and cost of high-throughput genome sequencing methods are leading to a rapid increase in not only the number of plant species sequenced, but also in the number of accessions sequenced within many species (Della Coletta *et al.* 2021; Jayakodi *et al.* 2021). Recently, high-quality genome sequences and annotations have been released

Received on 20 October 2023; accepted on 08 March 2024
Published by Oxford University Press on behalf of The Genetics Society of America 2024.

for the 26 maize NAM founder inbreds (Hufford et al. 2021), facilitating comprehensive comparative pan-genomic studies (Cagirici et al. 2022; Lovell et al. 2022; Thatcher et al. 2023). Annotation of the maize NAM founder inbred gene sequences was performed by applying state-of-the-art gene annotation methods using ab initio predictions and evidence-based predictions, including the use of transcriptomic data that were generated for all the NAM founder inbreds (Hufford et al. 2021; Li et al. 2022). Using the generated pan-genomic and pan-transcriptomic data for the NAM founder inbreds, a previous pan-genome coexpression network study showed substantial variation beyond the single reference genome and connected trait-specific genes with the pan-transcriptomic data (Cagirici et al. 2022), supporting other pan-transcriptome findings (Hirsch et al. 2014). These results suggest that the increase in the number of assembled pan-genomes and different types of available omics datasets will offer both opportunities and challenges for comparative pan-genomic studies.

Protein–protein interactions (PPIs) provide important insights into gene function and are considered to be a reliable indicator of functional associations (Wang et al. 2022). For proteins of interest, identification of interacting partners can elucidate the molecular basis for associated traits, such as sugar transport, phytohormone signaling, and flowering time (Garg et al. 2022; Zahn et al. 2023), or inform on possible strategies for trait improvement, such as plant development (Wang and Wang 2022). On the other hand, PPI networks can be used to identify novel regulatory interactions for targeted validation of protein function in complex signaling pathways, such as response to phytohormones and pathogen resistance (Jones et al. 2014; Altmann et al. 2020). Furthermore, integrative analyses of different multiomics datasets, including PPI and coexpression networks, can be used to dissect complex biological systems to identify target genes for crop improvement, such as flowering time (De Bodt et al. 2012; Han et al. 2023). Despite the large amount of experimental PPI data across different species, conditions, and organs (McWhite et al. 2020), the complexity and cost of high-throughput PPI discovery remains a challenge for interspecies and intraspecies pan-interactome analyses (Smits and Vermeulen 2016). In the absence of experimental PPI data, methods for genome-scale prediction of PPI network, such as the STRING database, offer a fast and scalable solution for generating predicted PPI networks from protein sequences alone (Szklarczyk et al. 2021). The STRING database contains experimental and predicted protein–protein interactions for physical and functional associations. These interactions were curated from computational predictions, knowledge transfers between organisms, high-throughput lab experiments, conserved coexpression data, automated text mining, and existing information in other databases, covering over 67 million proteins for more than 14,000 organisms (Szklarczyk et al. 2021). Using this information, it is possible to make reliable inferences and predictions of PPI networks. This not only advances our understanding of specific interacting proteins and individual PPI networks but also facilitates the comparisons between multispecies pan-interactomes.

In this study, we developed a framework for generating informative predicted pan-interactomes, using the established STRING database PPI prediction workflow (Szklarczyk et al. 2021), based on a selection of genomes lacking experimental PPI data. A variety of bioinformatics approaches have been developed and applied to the prioritization of gene candidates, including gene expression profiling (Woodhouse, Sen, et al. 2021), gene coexpression and PPI network analyses (Liu et al. 2019), and examination of relevant gene ontology (GO) terms (Almeida-Silva and Venancio 2022). We show that by using a PPI network clustering algorithm, we can generate simplified and informative clustered PPI networks that improve the overall interpretability of the predicted interactomes. Using this framework, we were able to create and annotate the clustered genome-, core-, and pan-interactome networks with information such as GO term enrichment-based functional annotations, coexpression data, and gene description annotations. Furthermore, we show that using GO enrichment analyses for cluster functional annotation can be leveraged for studying biological processes, predicting the function of proteins of interest, and prioritizing putative trait-associated gene candidates. While our pan-interactome analysis focuses on the recently assembled 26 NAM founder inbred genomes, our proposed framework can be extended to other pan-genomes, requiring only the annotated pan-gene mappings and protein sequences of each genome.

Materials and methods

Generating the predicted maize NAM-interactomes

The protein sequences of the latest B73 (RefGen_v5) reference genome and the 25 NAM founder genomes were obtained from MaizeGDB (Hufford et al. 2021; Woodhouse, Cannon, et al. 2021). For these genomes, protein sequences of the canonical transcripts were chosen based on domain coverage, protein length, and their similarity to assembled transcripts, representing a standard or reference version of a gene's structure (Hufford et al. 2021). The protein sequences of the canonical gene models were submitted to the STRING database for PPI network predictions (Szklarczyk et al. 2021). All predicted physical interactions were then used to construct and analyze the 26 individual maize NAM genome-interactomes. A list of all the generated STRING database accessions, including links to the STRING database download page, is available (Supplementary Table 1). While all the individual NAM genome-interactomes were processed and analyzed similarly, the B73-interactome was used as the representative genome-interactome in subsequent analyses.

Generating the pan- and core-interactomes

The pan-interactome network was created by mapping the protein IDs of the individual maize NAM genome-interactomes to the annotated MaizeGDB unified pan-gene IDs (Hufford et al. 2021). Only pan-PPIs, defined as pairs of interacting proteins that were both successfully mapped to a unified pan-gene ID, were included in the generation of the pan-interactome. The number of unique pan-PPIs in each of the 26 individual maize NAM genome-interactomes were counted, keeping all pan-PPIs that occurred in more than one genome-interactome to generate the final pan-interactome. The core-interactome, a subset of the pan-interactome, was created by keeping all pan-PPIs that were found in all 26 individual NAM genome-interactomes. Note that the protein IDs in the core- and pan-interactomes are based on the unified MaizeGDB pan-gene ID annotation, while the protein IDs in the individual NAM genome-interactomes are based on the genome-specific canonical gene IDs. All network graph figures were made with either Cytoscape (v3.10.1) (Shannon et al. 2003) or the Python NetworkX package (v3.1) (Hagberg et al. 2008).

Clustering and analysis of the genome-, core-, and pan-interactomes

A PPI network clustering approach was applied to the 26 individual genome-interactomes and to the core- and pan-interactomes to improve network interpretability. The interactomes were

clustered to identify densely connected PPIs using the overlapping graph clustering algorithm ClusterONE (v1.0, using the standalone Java application with default parameters) (Nepusz *et al.* 2012). Clusters were filtered based on a *P*-value < 0.1 to retain a larger number of clusters (Wisecaver *et al.* 2017). Because ClusterONE was designed for detection of protein complexes, we compared experimentally derived protein complexes with ClusterONE clusters of STRING-db PPI predictions (McWhite *et al.* 2020). The Jaccard similarity index score, based on the PyWGCNA method (Langfelder and Horvath 2008; Rezaie *et al.* 2023), was used to calculate overlap between ClusterONE clusters and experimental complexes that were detected using 4 different thresholds and mapped to the maize B73 (RefGen_v3) gene IDs (STRING-db accession STRG0A42DRC) using assigned eggNOG IDs (Huerta-Cepas *et al.* 2019). Additionally, we generated a cluster similarity network by comparing cluster pan-PPI overlap across the 26 individual genome-, core-, and pan-interactome clusters. After mapping the protein IDs to the annotated MaizeGDB unified pan-gene IDs, all clusters were compared based on overlap between cluster pan-PPIs members using the PyWGCNA method for calculating the Jaccard similarity index (Langfelder and Horvath 2008; Rezaie *et al.* 2023). The cluster similarity network was constructed by connecting clusters with a Jaccard similarity index >0.5.

Analysis of GO enrichment in PPI clusters

The GO-basic ontology was download from the GO consortium website and used for annotating the GO terms (2023-01-01 release) (Ashburner *et al.* 2000; Carbon and Mungall 2018; Gene Ontology Consortium *et al.* 2023). The GO annotations that were used for the subsequent GO term enrichment analyses were predicted using the PANNZER2 webserver (Törönen *et al.* 2018), using the same protein sequences that were submitted to STRING-db. To generate the GO annotation for the pan-genes, we combined all unique GO terms associated with each pan-gene from all available GO annotations. Analysis of GO enrichment was conducted using the Python package GOATOOLS (v1.3.1) (Klopfenstein *et al.* 2018). GO terms were considered to be enriched if the false discovery rate (FDR)-adjusted *P*-value was smaller than 0.05.

Coexpression analysis of interactome clusters

The complete quantified RNA-Seq data, i.e. fragments per kilobase of transcript per million mapped reads (FPKM) values, across 20 tissues for each of the 26 maize pan-genomes were obtained from the CyVerse Commons shared repository submitted by MaizeGDB (Hufford *et al.* 2021; Woodhouse, Cannon, *et al.* 2021). Each PPI cluster was processed using the Pearson's correlation coefficient (PCC) (SciPy v1.11.2) (Virtanen *et al.* 2020) for each pair of cluster genes, assigning genes to be coexpressed when PCC >0.9 and *P*-value < 0.05. The coexpression results for each cluster were used to supplement the PPI edges with coexpressed edges and edges that had both PPI and coexpression data. In the case of the core- and pan-interactome clusters, pan-genes were considered to be coexpressed if significant coexpression was observed in at least one of the individual NAM genome transcriptomes.

Generating gene descriptions from the *Arabidopsis thaliana* top DIAMOND hit

To provide additional informative gene description annotations, we have included the latest *A. thaliana* Araport11 (TAIR_Data_20220331) gene functional descriptions (Berardini *et al.* 2022). The best DIAMOND hit (v2.1.8.162) (Buchfink *et al.* 2021) between a given NAM founder inbred genome and *A. thaliana* was used to annotate the custom protein sequences included in the pan-interactome analysis. Pan-genes were annotated by selecting the NAM founder inbred gene with the longest annotation, including the reference gene ID from which the annotation was obtained.

A standalone Python Dash web application for accessing the annotated cluster data

To facilitate access to the generated data, we developed a standalone Python Dash web application (v2.13.0). The dash application takes one of two user inputs: (1) genes or (2) GO terms of interest. Based on the selected input, the web application identifies all relevant genome-, core-, and pan-interactome clusters containing either the genes or the enriched GO terms of interest. The cluster similarity network was used to identify all connected component groups of overlapping clusters based on a Jaccard similarity index score >0.5. The interface provides 4 output tabs that are updated based on the cluster selected: (1) a table containing all the enriched GO terms for the relevant clusters identified based on the user input, (2) a network graph showing the predicted PPIs and coexpression data between cluster members, (3) a table of the gene description annotation for cluster members based on protein sequence similarity to *A. thaliana* genes, and (4) a table containing all the enriched GO terms for similar clusters. The standalone web application and detailed installation instructions are available online at https://github.com/eporetsky/PanPPI/.

Results

Comparison of the functional annotation of the clustered interactomes

A protein interactome network for a given genome assembly is constructed using all the proteins present in that assembly. However, comparing single genome-based interactomes to interactomes of other genomes poses challenges. Using B73 as an example, not all proteins are included in the pan-protein set: 95.4% of proteins and 93.5% of pan-PPIs in the B73-interactome made it to the pan-protein set (Supplementary Table 2). Analysis of the number of shared pan-PPIs showed that a substantial number are either shared across more than 25 founder inbreds or are found in only five or fewer founder inbreds, with few found in between (Supplementary Fig. 1). Because the initial B73-, core-, and pan-interactomes yielded a "hairball"-like network that was not readily interpretable (Supplementary Fig. 2), ClusterONE was used to generate clustered interactomes from the individual 26 NAM genomes-, core-, and pan-interactomes to improve network interpretability (Fig. 1). We show that 338, 183, and 240, clusters were observed in the B73-, core-, and pan-interactomes, respectively (Fig. 2a). Among these clusters, a total of 3,633, 3,025, and 7,841, unique proteins were identified in the B73-, core-, and pan-interactomes, respectively (Fig. 2b and Supplementary Table 3). To assess the performance of ClusterONE to detect protein complexes, we compared our results with experimentally derived protein complexes obtained using four different thresholds (McWhite *et al.* 2020). We found that the number of predicted PPIs per complex increases with complex size (Supplementary Fig. 3a). On other hand, comparison of the overlap between experimental complexes and ClusterONE clusters showed that approximately 60% of complexes overlapped with clusters at a low Jaccard similarity index score cutoff of 0.1, but that the percent of overlapping clusters decreased as the Jaccard similarity score cutoff increased above 0.1 (Supplementary Fig. 3b). Treating the

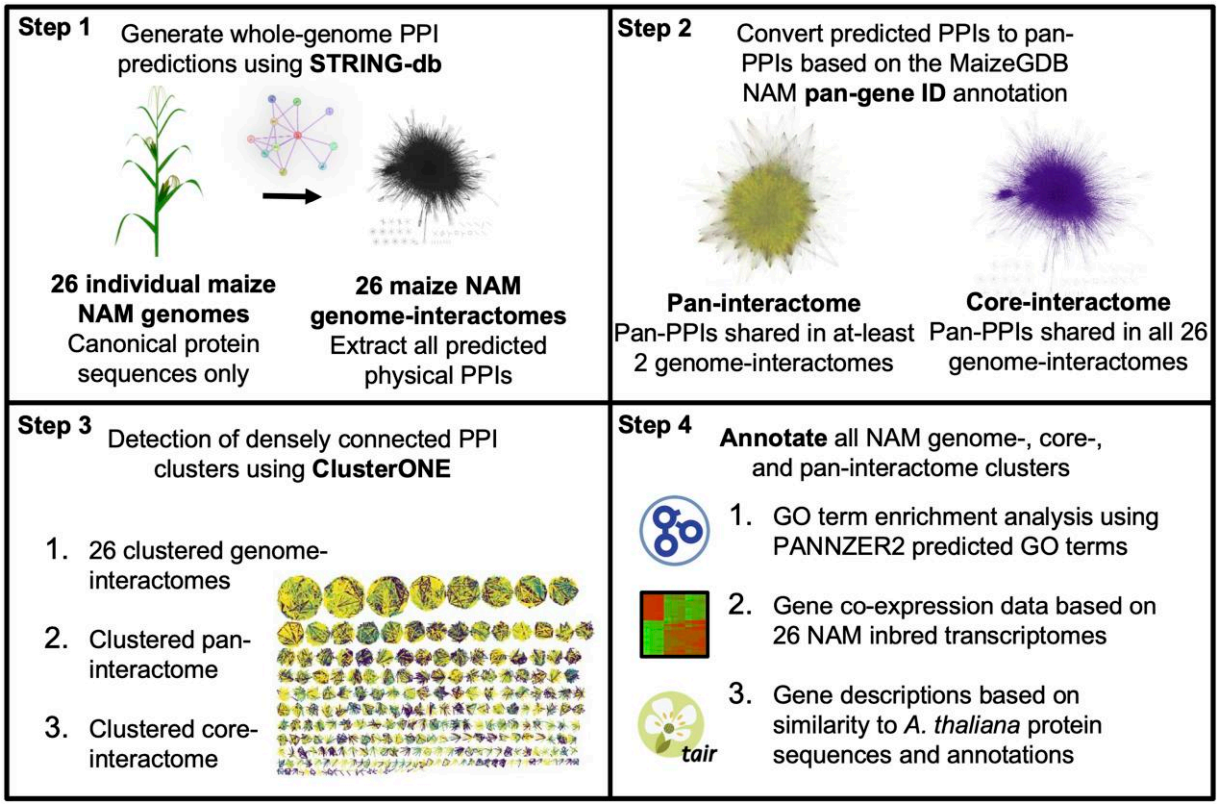


Fig. 1. Overview of the pan-genome analysis framework used for generating the annotated clustered maize NAM-, core-, and pan-interactomes. The workflow consists of 4 steps. First, STRING-db is used to generate the predicted PPI data for the 26 individual maize NAM genome-interactomes. Second, the MaizeGDB pan-gene annotation for the NAM founder inbreds is used to generate the core- and pan-interactomes. Third, ClusterONE is used to detect densely connected PPI clusters in the genome-, core-, and pan-interactomes. Finally, all the genome-, core-, and pan-interactome clusters are annotated using GO term enrichment analysis, gene coexpression data, and gene descriptions based on protein sequence similarity with *Arabidopsis thaliana* genes.

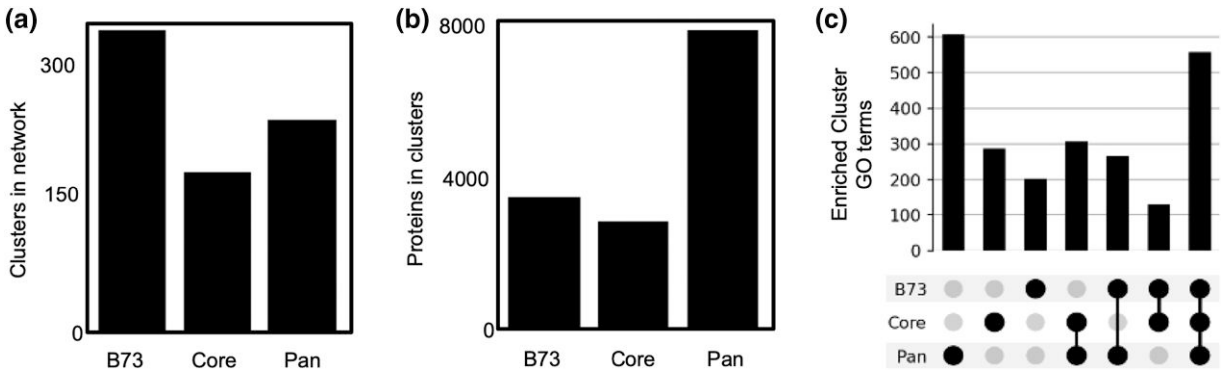


Fig. 2. Generation and functional annotation of the clustered interactomes. a) The number of clusters of densely connected PPIs in the predicted B73-, core-, and pan-interactomes detected by ClusterONE. b) The number of unique proteins found in the B73-, core-, and pan-interactome clusters. c) An upset plot showing the overlap of all significantly enriched GO terms (FDR-adjusted *P*-value < 0.05) found in an enrichment analysis of the B73-, core-, and pan-interactome clusters.

ClusterONE clusters as functionally associated groups of proteins, we further analyzed them using GO term enrichment analysis (Fig. 1) (Nepusz *et al.* 2012). An upset-plot comparison of the significantly enriched GO terms across the clusters of the B73-, core-, and pan-interactomes showed that many enriched GO terms were unique to each interactome, with many others shared by different interactome combinations (Fig. 2c). A higher degree of overlap between unique GO terms has been observed between the 26 individual clustered genome-interactomes (Supplementary Fig. 4). In addition to the clustered interactomes, we generated a cluster similarity network for the individual NAM-, core-, and pan-interactome clusters, based on pan-PPI overlap using a Jaccard similarity index score <0.5 (Supplementary Fig. 5). When comparing the different groups of overlapping clusters formed, we observe substantial differences in group sizes and the number of enriched biological process (BP) GO terms (Supplementary Fig. 5). In particular, in many groups of overlapping clusters, we find clusters with and without enriched BP GO terms, enabling

the use of additional functional annotation information from similar clusters (Supplementary Fig. 5).

Comparison of the functional annotation of the core- and pan-clustered interactomes

To compare the clustered core- and pan-interactomes, we generated a network graph showing each individual cluster, and colored the edges based on the number of times the pan-PPIs were observed across the different genome-interactomes (Fig. 3a and b). While comparing the 15 most significantly enriched BP GO terms in the core- and pan-interactome clusters, we observed a difference between larger (≥50 members) and smaller (<50 members) clusters (Supplementary Table 4). Larger clusters were primarily represented by general biological processes, such as translation, DNA replication, mRNA splicing, and rRNA processing (Supplementary Table 4), and had a large number of overlapping GO terms (Supplementary Table 4). On the other hand, the smaller core- and pan-interactome clusters had an overlap of 4 GO terms, involved in protein catabolism, transcription regulation, carbohydrate transport, and abscisic acid signaling (Fig. 3c and d, and Supplementary Table 4). The lists of enriched GO terms also show that while many enriched GO terms in the core-interactome are involved in general biological processes, such as auxin signaling, cell cycle, photosynthesis, and mRNA processing (Fig. 3c and Supplementary Table 4), the enriched GO terms in the pan-interactome represent more specialized biological processes, such as isoprenoid, arginine, and plastoquinone biosynthesis, methylation, jasmonic acid signaling, and signal transduction (Fig. 3d and Supplementary Table 4).

Integrating gene coexpression evidence with the clustered pan- and core-interactomes

Gene coexpression networks have been extensively used to derive indirect evidence for functional association (Wisecaver *et al.* 2017; Poretsky and Huffaker 2020; Cagirici *et al.* 2022) and can be integrated with different network types to enhance gene function prediction (Fig. 4a) (Han *et al.* 2023). By combining predicted PPIs with coexpression data, we were able to show that the number of clusters with any coexpression evidence was 47% for the B73-interactome clusters, 89% for the core-interactome clusters, and 97% for the pan-interactome clusters (Fig. 4b). When considering the total number of edges with only PPI evidence, coexpression evidence, or both, we observed that the number of PPI edges in the B73- and pan-interactomes were higher than both the coexpression-only edges and PPI with coexpression edges (Fig. 4c). In the core-interactome clusters, the number of coexpression-only edges was the highest (Fig. 4c). For the average number of edges per node with only PPI evidence, coexpression evidence, or both, the B73- and pan-interactome clusters have more PPI edges than coexpressed edges and both PPI and coexpressed edges (Fig. 4d). The B73 clusters have the lowest average coexpression-only edges per node (Fig. 4d). For the core-interactome clusters, the averages are similar across the 3 cases (Fig. 4d).

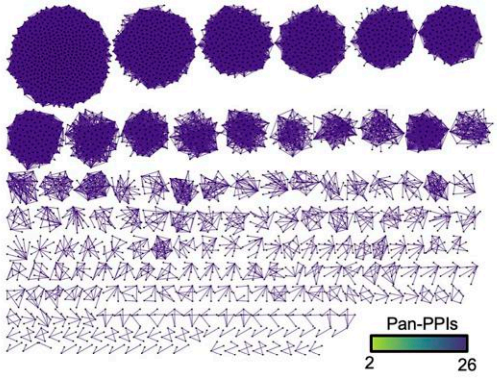
Leveraging functional enrichment of PPI clusters for putative gene function prediction

Physical interaction between proteins is considered to be a reliable indicator of functional association (Schwikowski *et al.* 2000). Based on the assumption that the PPI cluster members are functionally associated, we considered inferring gene function using the functional annotations for each cluster. In the case of the B73-interactome clusters, less than 60% of the clusters had at least one GO term enriched for BP and molecular function (MF), and less than 40% of the clusters had at least one GO term enriched for cellular component (CC) (Fig. 5a). A similar pattern was observed across the cluster NAM-interactomes for the enriched BP, MF, and CC GO terms, respectively (Supplementary Fig. 6a–c). In the core- and pan-interactome clusters, it was close to 80% for the BP- and MF-enriched GO terms, and approximately 50% for CC-enriched GO terms (Fig. 5a). Our framework allows users to retrieve useful functional information and infer putative gene functions based on clusters with relevant enriched GO term annotations. As an example, we considered clusters enriched for GO terms related to flowering time, an important agronomic trait with a relatively well understood genetic basis (Buckler *et al.* 2009; Dong *et al.* 2012). We searched for clusters enriched for GO terms with descriptions containing the keywords "photoperiodism" and "flowering" (GO:0048574, GO:2000028, GO:0048573, GO:0048578, GO:0048577, GO:0048579, GO:0048587, and GO:0048586) in the B73-, core-, and pan-interactome clusters and identified 3, 1, and 10 clusters, respectively (Fig. 5b and Supplementary Table 5). As an example, we selected cluster B73_184 that was enriched for the GO term GO:0048579 (negative regulation of long-day photoperiodism, flowering) (Fig. 5c and Supplementary Table 6). The B73_184 cluster contains Zm00001eb380460, a maize homolog of the core regulator of flowering time *CONSTANS* (*CO*) in *A. thaliana*, named *CONSTANS OF ZEA MAYS1* (*CONZ1*) (Fig. 5c) (Miller *et al.* 2008). Other cluster members, including four *NUCLEAR FACTOR-Y* (*NF-Y*) genes and one *BOI-RELATED GENE* (*BRG*) gene have been shown to regulate flowering time through a physical interaction with *CO* (Fig. 5c) (Nguyen *et al.* 2015; Myers and Holt 2018). Additionally, we find that *CONZ1* is both coexpressed and predicted to interact with Zm00001eb095880, member of the C2H2-like zinc finger transcription factor family, a gene family highly involved in transcriptional regulation of flowering induction and development (Fig. 5c) (Lyu and Cao 2018).

Leveraging functional enrichment of PPI clusters for prioritization of candidate genes

A multiomic study of maize development predicted 2,651 maize genes to be associated with flowering time regulation (Han *et al.* 2023). Of the predicted 2,651 genes, 20 genes were validated to alter flowering time in maize using knockout alleles generated through CRISPR–Cas9-mediated gene editing (Supplementary Table 7) (Han *et al.* 2023). Considering the 20 validated genes as candidate genes for our workflow, our analysis showed that 5, 4, and 7 clusters contained at least one of these candidate genes in the B73-, core-, and pan-interactome clusters, respectively (Fig. 6a). Among these clusters, we found a number of clusters that were enriched for GO terms with relevance to flowering, such as shoot development (GO:0010016, GO:2000032, and GO:0080006), flower development (GO:0048437, GO:0048444, and GO:0048455), and flowering time (GO:0048573, GO:0048574, GO:0048578, and GO:0010228), representing 7 of the 20 candidate genes (Fig. 6b and Supplementary Table 8). As an example, we selected cluster pan_503 that was enriched for the GO term GO:0010228 (vegetative to reproductive phase transition of meristem). Cluster pan_503 contains pan_gene_18303, associated with the candidate gene Zm00001eb155150, that shares sequence similarity to the *A. thaliana FLOWERING LOCUS VE* (*FVE*) gene (Fig. 6c and Supplementary Table 9). Additionally, the candidate gene, Zm00001eb155150, was found to be coexpressed with 12 other cluster members, in addition to two predicted pan-PPIs (Fig. 6c). Based on the descriptions of the top DIAMOND hits to *A. thaliana* genes, the cluster contains 9 *FVE*-similar genes, a

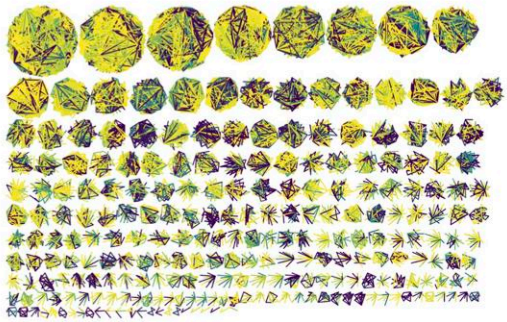
(a) Core-interactome clusters



(c)

Cluster	Term	p-val	Description
core_73	GO:0031145	5.84E-39	anaphase-promoting complex-dependent catabolic process
core_232	GO:0009734	2.13E-36	auxin-activated signaling pathway
core_171	GO:1902600	9.53E-33	proton transmembrane transport
core_515	GO:0006260	2.68E-23	DNA replication
core_81	GO:0032509	7.98E-21	endosome transport via multivesicular body sorting pathway
core_116	GO:0015979	1.98E-19	photosynthesis
core_408	GO:0045944	6.90E-19	positive regulation of transcription by RNA polymerase II
core_397	GO:0034219	4.05E-18	carbohydrate transmembrane transport
core_370	GO:0006621	5.69E-18	protein retention in ER lumen
core_17	GO:0051560	5.69E-17	mitochondrial calcium ion homeostasis
core_292	GO:0009765	6.41E-17	photosynthesis, light harvesting
core_302	GO:0006086	8.35E-17	acetyl-CoA biosynthetic process from pyruvate
core_324	GO:0009738	1.08E-16	abscisic acid-activated signaling pathway
core_346	GO:0032509	1.19E-16	endosome transport via multivesicular body sorting pathway
core_305	GO:0030163	8.39E-16	protein catabolic process

(b) Pan-interactome clusters



(d)

Cluster	Term	p-val	Description
pan_542	GO:0045944	1.42E-51	positive regulation of transcription by RNA polymerase II
pan_726	GO:0045944	2.50E-50	positive regulation of transcription by RNA polymerase II
pan_182	GO:0043414	5.27E-42	macromolecule methylation
pan_694	GO:0009738	9.34E-38	abscisic acid-activated signaling pathway
pan_444	GO:0008299	1.95E-36	isoprenoid biosynthetic process
pan_886	GO:0006032	7.81E-33	chitin catabolic process
pan_947	GO:0010236	3.66E-32	plastoquinone biosynthetic process
pan_157	GO:0045944	1.90E-27	positive regulation of transcription by RNA polymerase II
pan_120	GO:0006526	5.36E-27	arginine biosynthetic process
pan_104	GO:0006508	8.00E-26	proteolysis
pan_963	GO:0009738	1.32E-25	abscisic acid-activated signaling pathway
pan_174	GO:0030163	9.11E-25	protein catabolic process
pan_1027	GO:0009867	2.46E-24	jasmonic acid mediated signaling pathway
pan_908	GO:0034219	1.48E-23	carbohydrate transmembrane transport
pan_395	GO:0000160	2.30E-23	phosphorelay signal transduction system

Fig. 3. Comparison of the core- and pan-interactome clusters and functional enrichments. a and b) Cytoscape network graphs of the core- and pan-interactome clusters, respectively. Edge density represents the number of times pan-PPIs are shared across multiple genome-interactomes. c and d) The top 15 most enriched BP GO terms among the core- and pan-interactome clusters that have less than 50 cluster members, respectively. Enrichment P-values were calculated using a hypergeometric test and adjusted using the Bonferroni method.

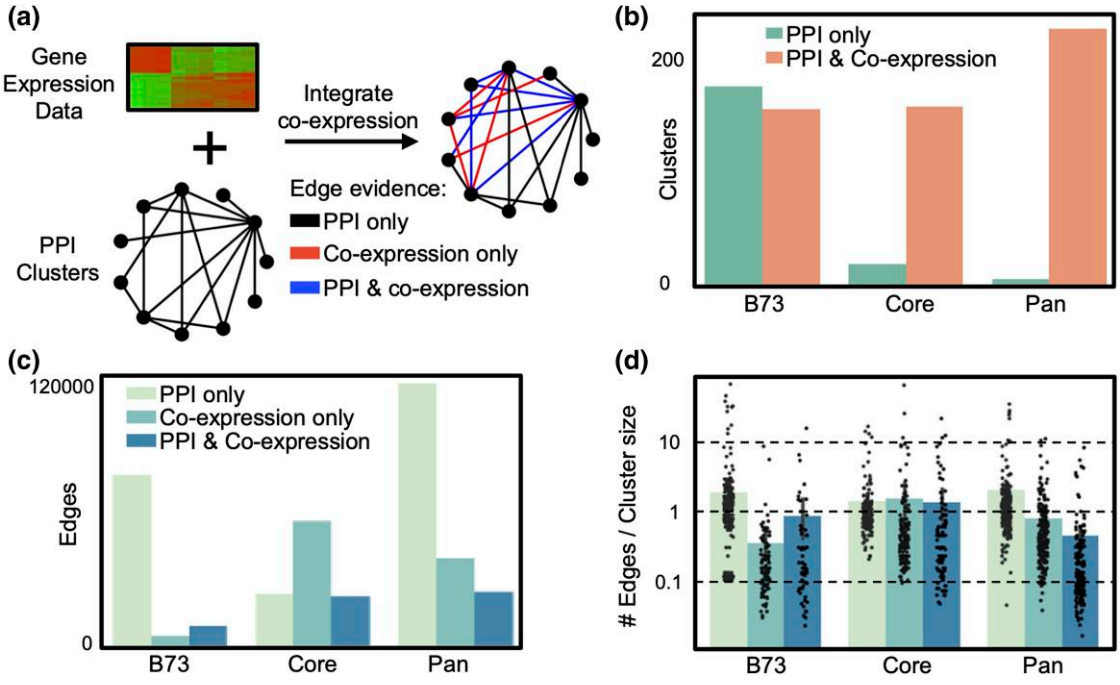


Fig. 4. Integration of PPI cluster networks with gene coexpression data. a) Outline of the PPI cluster and gene coexpression data integration. b) Total number of clusters in the B73-, core-, and pan-interactomes with only predicted PPI edges or clusters containing both PPI and gene coexpression edges. c) Total number of edges in all clusters of the B73-, core-, and pan-interactomes with PPI, coexpression, or both annotations. d) Number of edges normalized by cluster size for all B73-, core-, and pan-interactome clusters with PPI, coexpression, or both annotations.

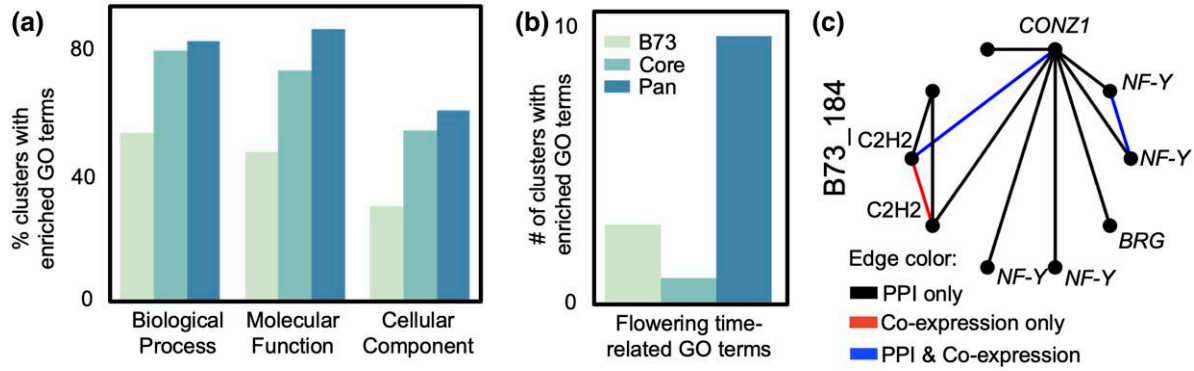


Fig. 5. Using functional enrichment of interactome clusters to search for relevant clusters. a) Comparison of the percent of B73-, core-, and pan-interactome clusters with one or more enriched GO terms in the 3 GO domains. b) Number of unique clusters enriched for flowering time-related GO terms that contain "flowering" and "photoperiodism" in their GO term descriptions (GO:0048574, GO:2000028, GO:0048573, GO:0048578, GO:0048577, GO:0048579, GO:0048587, and GO:0048586). c) The B73_184 cluster is enriched for the GO term GO:0048579 (negative regulation of long-day photoperiodism, flowering). Labeled nodes represent gene descriptions related to flowering time based on the gene description annotation, including *CONSTANS OF MAIZE1* (*CONZ1*), *NF-Y*, *BRG*, and C2H2-like zinc finger transcription factor family genes.

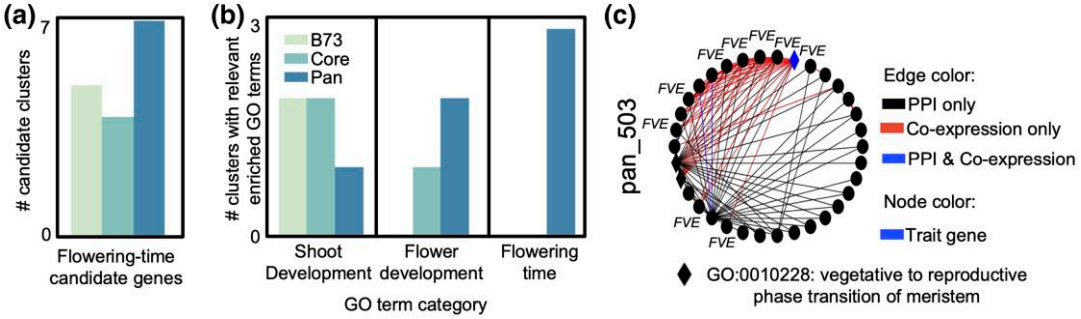


Fig. 6. Identifying relevant clusters from omics-related flowering time candidate genes. a) A list of 20 flowering time candidate genes, obtained from Han *et al.* (2023), were used to find the number of associated clusters in the B73-, core-, and pan-interactome. b) Number of clusters containing relevant enriched GO terms split into three categories: (1) shoot development, consisting of GO:0010016 (shoot system morphogenesis), GO:2000032 (regulation of secondary shoot formation), and GO:0080006 (internode patterning), (2) flower development, consisting of GO:0048437 (floral organ development), GO:0048444 (floral organ morphogenesis), and GO:0048455 (stamen formation), and (3) flowering time, consisting of GO:0048573 (photoperiodism, flowering), GO:0048574 (long-day photoperiodism, flowering), GO:0048578 (positive regulation of long-day photoperiodism, flowering), and GO:0010228 (vegetative to reproductive phase transition of meristem). c) Graph of the pan_503 cluster containing the flowering time candidate gene, Zm00001eb155150 (pan_gene_18303), 1 of 8 *FVE* genes. Nodes annotated with the GO:0010228 term are diamond shaped.

FLOWERING LOCUS D (*FLD*)-similar gene, 7 histone deacetylase genes, 4 adenine-thymine (AT)-hook motif containing genes, a ubiquitin-protein ligase, and 12 unannotated genes (Fig. 6c and Supplementary Table 9). In *A. thaliana*, *FLD* and *FVE* are part of the flowering autonomous pathway that restricts *FLOWERING LOCUS C* (*FLC*) expression to promote transition to flowering (Ausín *et al.* 2004). Histone deacetylation by histone deacetylases in the *FLC* chromatin was shown to regulate flowering by down-regulating *FLC* expression (He *et al.* 2003), possibly through physical interaction between FLD and histone deacetylases (Yu *et al.* 2011). The 4 AT-hook genes were most similar to *AT-HOOK MOTIF NUCLEAR-LOCALIZED PROTEIN22* that was shown to regulate flowering time by promoting acetylation and methylation in the *FLOWERING LOCUS T* chromatin (Yun *et al.* 2012). The ubiquitin-protein ligase was most similar to *HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE GENES1*, shown to regulate flowering through physical interaction with CO to promote *FLC* expression (Lazaro *et al.* 2012). Of the 34 clusters genes, only the 4 AT-hook genes were annotated with the GO:0010228 term (Supplementary Table 9).

As a second example, we examined a list of trait-specific gene candidates from an association mapping study conducted in the maize NAM mapping population for diverse metabolomic traits (Wallace *et al.* 2014; Wang *et al.* 2022). Focusing on the trait-specific genes found in proximity to loci associated with 2 metabolomic traits, the first principal component (PC1) and PC2 traits of the metabolite data (Supplementary Table 7), we searched for relevant enriched GO terms in the B73-, core-, and pan-interactome clusters. We first looked at the PCCs of the PC1 and PC2 traits with each of the analyzed metabolites and found that PC1 had the highest correlation with glutamate, chlorophyll A, and malate levels (PCCs of 0.74, 0.62, and 0.61, respectively), with glutamate being the precursor for chlorophyll (Tanaka and Tanaka 2006), while PC2 was most highly correlated with glucose, starch, and fructose levels (PCCs of 0.71, 0.7, and 0.7, respectively) (Fig. 7a). More than 8 clusters in each of the B73-, core-, and pan-interactomes were found to contain PC1 and PC2 trait-specific gene candidates (Fig. 7b). Within the PC1 trait-specific clusters, we found 7 clusters to be enriched for a GO term associated chlorophyll catabolism (GO:0015996) and within the PC2 trait-specific clusters we found only one cluster to be associated with UDP-xylose transmembrane transport (GO:0015790) (Fig. 7c and Supplementary Table 10). Cluster network graphs show that the trait-specific genes for both the PC1 and PC2 traits were annotated with the

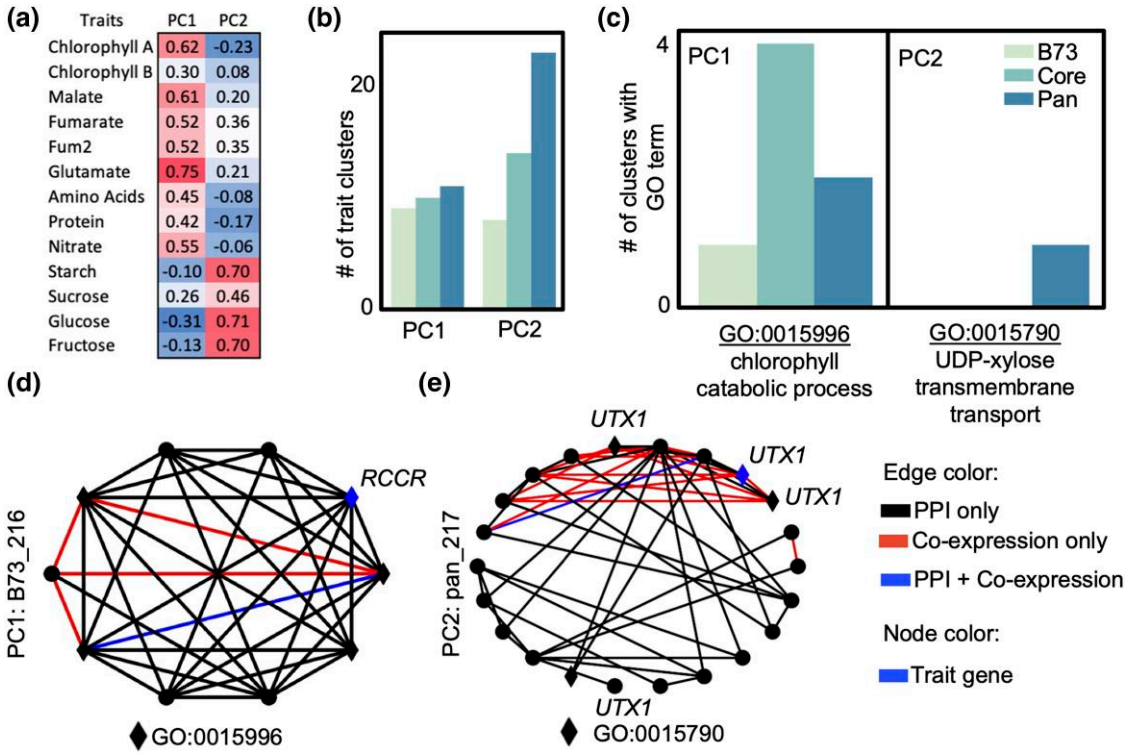


Fig. 7. Identifying relevant clusters for association mapping-related metabolomic candidate genes. a) Correlation between the metabolite PC1 and PC2 traits with the individual metabolites, measured across the NAM mapping population in Wallace *et al.* (2014). b) Number of identified PC1 and PC2 trait-specific clusters in the B73, core-, and pan-interactomes. c) Number of clusters containing the GO term GO:0015996 (chlorophyll catabolic process) and GO:0015790 (UDP-xylose transmembrane transport) associated with the PC1 and PC2 trait-specific clusters, respectively. d and e) Cluster graphs of the PC1 trait B73_216 and PC2 trait pan_217 clusters with the labeled *A. thaliana*-based gene descriptions of the candidate genes *RCCR* and *UXT1*, respectively. Nodes annotated with specified GO terms are diamond shaped.

relevant GO term and that while the trait-specific gene in the PC2 had evidence for both PPI and coexpression with other cluster members, the PC1 trait-specific gene had only PPI evidence (Fig. 7d and e). A closer inspection of the trait-specific genes showed that, according to the top *A. thaliana* DIAMOND hit, the PC1 trait-specific gene (Zm00001eb027950) encodes *RED CHLOROPHYLL CATABOLITE REDUCTASE* (*RCCR*) involved in the chlorophyll breakdown pathway (Fig. 7d and Supplementary Table 11) (Sugishima *et al.* 2010). Six other cluster members are described as involved in chlorophyll degradation and catabolism (Supplementary Table 11). The PC2 trait-specific gene (pan_gene_6500), with 3 other cluster members, were similar to the *A. thaliana UDP-XYLOSE TRANSPORTER1* (*UXT1*), with *UXT1* mutants showing altered monosaccharide composition, including altered UDP-glucose levels (Fig. 7e and Supplementary Table 11) (Zhao *et al.* 2018). Additionally, pan_gene_6500 was found to be coexpressed with four other cluster members, in addition to the 2 predicted pan-PPIs (Fig. 7e). Two other pan_217 cluster members were described as aluminum activated malate transporters (Supplementary Table 11), with evidence showing malate content to be negatively correlated with starch and soluble sugars content (Centeno *et al.* 2011).

Discussion

Protein interaction networks are a reliable source for functional association prediction and are often used for understanding complex biological processes (Schwikowski *et al.* 2000; Wang *et al.* 2022). Although pan-genome analyses are becoming increasingly prevalent and useful, most plant species lack experimental PPI data beyond the reference species (McWhite *et al.* 2020). In this study, we generated predicted PPI networks for the 26 maize NAM founder inbreds which were used to generate the clustered genome-, core- and pan-interactomes. We show that in contrast to the small number of coexpressed pan-genes shared across the majority of maize NAM founder inbreds (Cagirici *et al.* 2022), a large number of PPIs were shared among the predicted genome PPI networks (Supplementary Fig. 1). Nonetheless, many PPIs were identified in only a few predicted genome PPI networks, suggesting a benefit for retaining the individual genome-interactomes for comparative pan-interactome studies (Supplementary Fig. 1). Due to the complexity and high interconnectedness of the predicted genome-, core-, and pan-interactomes (Supplementary Fig. 2), we show that PPI clustering improved interpretability and facilitated identification of putative groups of functionally associated proteins (Fig. 2a–c). We find that while STRING-db captures a substantial number of PPIs between experimentally derived complex members (Supplementary Fig. 3a), using ClusterONE on STRING-db predicted PPIs is more suitable for identifying densely connected PPI clusters than recovery of protein complexes (Supplementary Fig. 3b). Comparison of the enriched BP GO terms in the core- and pan-interactome clusters showed that clusters larger than 50 members are similarly enriched for general GO terms (Supplementary Table 4). On the other hand, smaller core-interactome clusters were more highly enriched for general BP GO terms and smaller pan-interactome clusters were more highly enriched for specialized BP GO terms (Fig. 3c and d), similar to the observation made in an *A. thaliana* pan-transcriptomic analysis (He and Maslov 2016). Furthermore, we observed that keeping the individual genome-interactome clusters, in addition to

the core- and pan-interactomes, substantially increases the number of unique enriched GO terms (Supplementary Fig. 4) and increases the diversity of functional annotations for comparative pan-interactome analyses (Supplementary Fig. 5).

Gene coexpression analyses have been extensively used to provide indirect evidence for functional association across species and conditions, facilitating the elucidation of different functional associations and biological pathways (Wisecaver *et al.* 2017). For this reason, we sought to integrate gene coexpression data from the NAM founder inbreds with the predicted PPI clusters to provide additional supporting information for functional association (Fig. 4a). Pan-genomes offer an opportunity to extend existing gene candidate prioritization approaches by integrating information from multiple genomes in relation to the studied trait. For example, pan-genome graph-based genetic mapping approaches have been able to identify novel trait-associated genetic markers that were missing from the traditional reference genome-based genetic mapping approaches but that were present in the pan-genome (Della Coletta *et al.* 2021). Furthermore, pan-genome coexpression analyses can be used to construct trait-specific coexpression pan-networks to identify groups of coregulated genes (Cagirici *et al.* 2022). Thus, integrating gene coexpression data with the PPI clusters adds a considerable number of new connections between interactome clusters members for almost half of the B73-interactome clusters and over 90% of the core- and pan-interactome clusters (Fig. 4b–d). In one example, the predicted PPI between *CONZ1* and a C2H2 transcription factor was supplemented with a coexpression interaction (Fig. 5c). Considering that coexpression between interacting proteins can coevolve to maintain stoichiometry among interacting partners, such information could indicate a biologically significant functional associations (Fraser *et al.* 2004; Piya *et al.* 2014). In two other examples, gene coexpression data provided additional connections between candidate genes and cluster members, as shown for the flowering time *FVE* candidate gene in cluster pan_503 and for the PC2 metabolomic trait *UTX1* candidate gene in cluster pan_217 (Figs. 6c and 7e). Despite not being connected by predicted PPIs, coexpression edges between candidate genes and other interactome cluster members are a useful indicator for functional association underlying a given trait of interest (Ficklin *et al.* 2010; De Bodt *et al.* 2012).

Protein interaction networks have been extensively used for the identification of functionally associated proteins and function prediction (Schwikowski *et al.* 2000; Szklarczyk *et al.* 2021; Wang *et al.* 2022), offering a better understanding of biological and molecular functions. In contrast to experimental PPI networks, the use of predicted PPI networks for network analysis and protein function prediction has been limited (Lin *et al.* 2011; Musungu *et al.* 2015). Nonetheless, until sufficient experimental PPI data are produced for individual genomes and pan-genomes, predicted PPI networks offer a promising opportunity for comparative interactome studies (Wang *et al.* 2022). The enrichment analysis of clustered interactomes allows researchers to search for GO terms of interest and to identify relevant enriched clusters. For example, when searching for clusters enriched for GO terms associated with flowering time regulation, we found cluster B73_184 to contain cluster members with both known and unknown roles in flowering time regulation, based on similarity to *A. thaliana* genes (Fig. 5c and Supplementary Table 6). Of the 10 B73_184 cluster members, only *CONZ1* and one of four *NF-Y* genes were annotated with the enriched flowering time GO:0048579 term (Supplementary Table 6), suggesting that the information from the annotated clusters can provide support for putative protein function predictions (Letovsky and Kasif 2003). Thus, researchers studying specific biological processes can search for relevant clusters and use the provided cluster annotations to predict functional associations and putative protein functions (Ficklin *et al.* 2010). Researchers can also search the clustered interactomes for clusters containing candidate genes of interest, such as obtained from omics-related and association mapping studies. In this case, clusters with trait-relevant enriched GO terms can be used to prioritize lists of candidate genes (Ficklin *et al.* 2010). In one example, 7 out of 20 verified flowering time candidate genes were present in clusters with relevant enriched GO terms (Fig. 6b and Supplementary Table 8), including the putative flowering time regulator *FVE* candidate gene in the pan_503 cluster (Fig. 6c) (Han *et al.* 2023). In a second example, we identified clusters enriched for chlorophyll catabolism and carbohydrate transport GO terms, relevant to the candidate genes obtained from an association mapping study using the PC1 and PC2 metabolomic traits, respectively (Fig. 7d and e) (Wallace *et al.* 2014). In both cases, the cluster annotations provide evidence for possible causal links between the candidate genes and the associated traits, in addition to providing useful information about the cluster members and their functional associations (Supplementary Tables 9 and 11). We anticipate that the results generated in this study will enable researchers in different fields, including biochemists, molecular biologists, and geneticists, to harness the annotated clusters to better understand interactions between genes, and for obtaining useful information and hypothesis generation.

Conclusion

A major advantage of our proposed pan-interactome analysis approach is the reliance on an established PPI prediction method, namely the STRING database, in generating the input data required for the analysis. For practical reasons, this means that our analysis workflow can be easily adapted to any set of genomes, including for interspecies and intraspecies analyses. To gain useful insights from the predicted PPI networks, we applied a PPI clustering algorithm, namely ClusterONE, to extract putative functionally meaningful PPI clusters, effectively disentangling the complex raw "hairball"-like PPI networks. By including the genome-interactomes, together with core- and pan-interactomes, we show that we capture substantially more functionally enriched clusters with unique GO term annotations. Additionally, our method allows the simple integration of supporting information such as gene coexpression and gene description annotations with the predicted interactomes, significantly increasing the breadth of genomic annotations that can be included in the pan-interactome analysis. Furthermore, we show that using functional enrichment to annotate PPI clusters can be used for putative protein function prediction and prioritization of trait-specific candidate gene sets. We anticipate that improved PPI prediction methods and gene function annotation (Odell *et al.* 2017) will further improve the annotation of the predicted PPI-interactome clusters.

Data availability

The files used in the preparation of the manuscript and the generated results have been submitted to figshare: https://doi.org/10.25387/g3.25301212. The source code for the Python implementation of the analysis workflow is available at github.com/eporetsky/PanPPI.

Supplemental material is available at G3 online.

Acknowledgments

This research was supported in part by an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA).

Funding

This research was supported by the U.S. Department of Agriculture, Agricultural Research Service, Project No. 2030-21000-056-00D through the Crop Improvement and Genetics Research Unit and Project No. 5030-21000-072-00D through the Corn Insects and Crop Genetics Research Unit. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Conflicts of interest

The author(s) declare no conflicts of interest.

Author contributions

EP: conceptualization, data curation, formal analysis, methodology, pipeline design, validation, visualization, writing—original draft preparation, writing—review and editing. HBC: conceptualization, data curation, formal analysis, methodology, pipeline design, validation, visualization, writing—original draft preparation, writing—review and editing. CMA: conceptualization, project administration, writing—review and editing. TZS: conceptualization, funding acquisition, project administration, supervision, writing—review and editing. All authors contributed to the article and approved the submitted version.

Literature cited

Almeida-Silva F, Venancio TM. 2022. *cageminer*: an R/Bioconductor package to prioritize candidate genes by integrating genome-wide association studies and gene coexpression networks. In silico Plants. 4(2):diac018. doi:10.1093/insilicoplants/diac018.

Altmann M, Altmann S, Rodriguez PA, Weller B, Elorduy Vergara L, Palme J, Marín-de La Rosa N, Sauer M, Wenig M, Villaécija-Aguilar JA, *et al.* 2020. Extensive signal integration by the phytohormone protein network. Nature. 583(7815):271–276. doi:10.1038/s41586-020-2460-0.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25(1):25–29. doi:10.1038/75556.

Ausín I, Alonso-Blanco C, Jarillo JA, Ruiz-García L, Martínez-Zapater JM. 2004. Regulation of flowering time by FVE, a retinoblastoma-associated protein. Nat Genet. 36(2):162–166. doi:10.1038/ng1295.

Berardini T, Reiser L, Huala E. 2022. TAIR functional annotation data [accessed 2023 Sep 12]. Available from https://zenodo.org/record/7843882.

Broekema RV, Bakker OB, Jonkers IH. 2020. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biol. 10(1):190221. doi:10.1098/rsob.190221.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 18(4):366–368. doi:10.1038/s41592-021-01101-x.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, *et al.* 2009. The genetic architecture of maize flowering time. Science. 325(5941):714–718. doi:10.1126/science.1174276.

Cagirici HB, Andorf CM, Sen TZ. 2022. Co-expression pan-network reveals genes involved in complex traits within maize pan-genome. BMC Plant Biol. 22(1):595. doi:10.1186/s12870-022-03985-z.

Centeno DC, Osorio S, Nunes-Nesi A, Bertolo ALF, Carneiro RT, Araújo WL, Steinhauser M-C, Michalska J, Rohrmann J, Geigenberger P, *et al.* 2011. Malate plays a crucial role in starch metabolism, ripening, and soluble solid content of tomato fruit and affects postharvest softening. Plant Cell. 23(1):162–184. doi:10.1105/tpc.109.072231.

De Bodt S, Hollunder J, Nelissen H, Meulemeester N, Inzé D. 2012. CORNET 2.0: integrating plant coexpression, protein–protein interactions, regulatory interactions, gene associations and functional annotations. New Phytologist. 195(3):707–720. doi:10.1111/j.1469-8137.2012.04184.x.

Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN. 2021. How the pan-genome is changing crop genomics and improvement. Genome Biol. 22(1):3. doi:10.1186/s13059-020-02224-8.

Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. 2012. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One. 7(8):e43450. doi:10.1371/journal.pone.0043450.

Ficklin SP, Luo F, Feltus FA. 2010. The association of multiple interacting genes with specific phenotypes in rice using gene coexpression networks. Plant Physiol. 154(1):13–24. doi:10.1104/pp.110.159459.

Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A. 101(24):9033–9038. doi:10.1073/pnas.0402591101.

Garg V, Reins J, Hackel A, Kühn C. 2022. Elucidation of the interactome of the sucrose transporter StSUT4: sucrose transport is connected to ethylene and calcium signalling. J Exp Bot. 73(22):7401–7416. doi:10.1093/jxb/erac378.

Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, *et al.* 2023. The gene ontology knowledgebase in 2023. Genetics. 224(1):iyad031. doi:10.1093/genetics/iyad031.

Carbon S, Mungall C. 2018. Gene ontology data archive [accessed 2023 Oct 6]. Available from https://zenodo.org/record/7504797.

Hagberg AA, Schult DA, Swart PJ. 2008. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena (CA). 11–15.

Han L, Zhong W, Qian J, Jin M, Tian P, Zhu W, Zhang H, Sun Y, FengJ-W, Liu X, *et al.* 2023. A multi-omics integrative network map of maize. Nat Genet. 55(1):144–153. doi:10.1038/s41588-022-01262-1.

He F, Maslov S. 2016. Pan- and core- network analysis of co-expression genes in a model plant. Sci Rep. 6(1):38956. doi:10.1038/srep38956.

He Y, Michaels SD, Amasino RM. 2003. Regulation of flowering time by histone acetylation in *Arabidopsis*. Science. 302(5651):1751–1754. doi:10.1126/science.1091109.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, *et al.* 2014. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 26(1):121–135. doi:10.1105/tpc.113.119982.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, *et al.* 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47(D1):D309–D314. doi:10.1093/nar/gky1085.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, *et al.* 2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 373(6555):655–662. doi:10.1126/science.abg5289.

Jayakodi M, Schreiber M, Stein N, Mascher M. 2021. Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Res. 28(1):dsaa030. doi:10.1093/dnares/dsaa030.

Jones AM, Xuan Y, Xu M, Wang R-S, Ho C-H, Lalonde S, You CH, Sardi MI, Parsa SA, Smith-Valle E, *et al.* 2014. Border control—a membrane-linked interactome of *Arabidopsis*. Science. 344(6185):711–716. doi:10.1126/science.1251358.

Kersey PJ. 2019. Plant genome sequences: past, present, future. Curr Opin Plant Biol. 48:1–8. doi:10.1016/j.pbi.2018.11.001.

Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, WeigelM, *et al.* 2018. GOATOOLS: a python library for gene ontology analyses. Sci Rep. 8(1):10872. doi:10.1038/s41598-018-28948-z.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9(1):559. doi:10.1186/1471-2105-9-559.

Lazaro A, Valverde F, Piñeiro M, Jarillo JA. 2012. The *Arabidopsis* E3 ubiquitin ligase HOS1 negatively regulates CONSTANS abundance in the photoperiodic control of flowering. Plant Cell. 24(3):982–999. doi:10.1105/tpc.110.081885.

Letovsky S, Kasif S. 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 19(suppl. 1):i197–i204. doi:10.1093/bioinformatics/btg1026.

Li J, Singh U, Bhandary P, Campbell J, Arendsee Z, Seetharam AS, Wurtele ES. 2022. Foster thy young: enhanced prediction of orphan genes in assembled genomes. Nucleic Acids Res. 50(7):e37. doi:10.1093/nar/gkab1238.

Lin M, Zhou X, Shen X, Mao C, Chen X. 2011. The predicted Arabidopsis interactome resource and network topology-based systems biology analyses. Plant Cell. 23(3):911–922. doi:10.1105/tpc.110.082529.

Liu W, Lin L, Zhang Z, Liu S, Gao K, Lv Y, Tao H, He H. 2019. Gene co-expression network analysis identifies trait-related modules in *Arabidopsis thaliana*. Planta. 249(5):1487–1501. doi:10.1007/s00425-019-03102-9.

Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J. 2022. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. eLife. 11:e78526. doi:10.7554/eLife.78526.

Lyu T, Cao J. 2018. Cys2/His2 zinc-finger proteins in transcriptional regulation of flower development. IJMS. 19(9):2589. doi:10.3390/ijms19092589.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, *et al.* 2009. Genetic properties of the maize nested association mapping population. Science. 325(5941):737–740. doi:10.1126/science.1174320.

McWhite CD, Papoulas O, Drew K, Cox RM, June V, Dong OX, Kwon T, Wan C, Salmi ML, Roux SJ, *et al.* 2020. A pan-plant protein complex map reveals deep conservation and novel assemblies. Cell. 181(2):460–474.e14. doi:10.1016/j.cell.2020.02.049.

Miller TA, Muslin EH, Dorweiler JE. 2008. A maize CONSTANS-like gene, conz1, exhibits distinct diurnal expression patterns in varied photoperiods. Planta. 227(6):1377–1388. doi:10.1007/s00425-008-0709-1.

Mural RV, Sun G, Grzybowski M, Tross MC, Jin H, Smith C, Newton L, Andorf CM, Woodhouse MR, Thompson AM, *et al.* 2022. Association mapping across a multitude of traits collected in diverse environments in maize. GigaScience. 11:giac080. doi:10.1093/gigascience/giac080.

Musungu B, Bhatnagar D, Brown RL, Fakhoury AM, Geisler M. 2015. A predicted protein interactome identifies conserved global networks and disease resistance subnetworks in maize. Front Genet. 6. doi:10.3389/fgene.2015.00201.

Myers ZA, Holt BF. 2018. NUCLEAR FACTOR-Y: still complex after all these years? Curr Opin Plant Biol. 45:96–102. doi:10.1016/j.pbi.2018.05.015.

Nepusz T, Yu H, Paccanaro A. 2012. Detecting overlapping protein complexes in protein–protein interaction networks. Nat Methods. 9(5):471–472. doi:10.1038/nmeth.1938.

Nguyen KT, Park J, Park E, Lee I, Choi G. 2015. The Arabidopsis RING domain protein BOI inhibits flowering via CO-dependent and CO-independent mechanisms. Mol Plant. 8(12):1725–1736. doi:10.1016/j.molp.2015.08.005.

Odell SG, Lazo GR, Woodhouse MR, Hane DL, Sen TZ. 2017. The art of curation at a biological database: principles and application. Curr Plant Biol. 11–12:2–11. doi:10.1016/j.cpb.2017.11.001.

Pan Q, Xu Y, Li K, Peng Y, Zhan W, Li W, Li L, Yan J. 2017. The genetic basis of plant architecture in 10 maize recombinant inbred line populations. Plant Physiol. 175(2):858–873. doi:10.1104/pp.17.00709.

Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z, Millard MJ, Gardner CAC, McMullen MD, Holland JB, Bradbury PJ, *et al.* 2014. The genetic architecture of maize height. Genetics. 196(4):1337–1356. doi:10.1534/genetics.113.159152.

Piya S, Shrestha SK, Binder B, Stewart CN, Hewezi T. 2014. Protein–protein interaction and gene co-expression maps of ARFs and Aux/IAAs in Arabidopsis. Front Plant Sci. 5. doi:10.3389/fpls.2014.00744.

Poretsky E, Huffaker A. 2020. MutRank: an R shiny web-application for exploratory targeted mutual rank-based coexpression analyses integrated with user-provided supporting information. PeerJ. 8:e10264. doi:10.7717/peerj.10264.

Rezaie N, Reese F, Mortazavi A. 2023. PyWGCNA: a Python package for weighted gene co-expression network analysis. Bioinformatics. 39(7):btad415. doi:10.1093/bioinformatics/btad415.

Schwikowski B, Uetz P, Fields S. 2000. A network of protein–protein interactions in yeast. Nat Biotechnol. 18(12):1257–1261. doi:10.1038/82360.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498–2504. doi:10.1101/gr.1239303.

Shi J, Tian Z, Lai J, Huang X. 2023. Plant pan-genomics and its applications. Mol Plant. 16(1):168–186. doi:10.1016/j.molp.2022.12.009.

Smits AH, Vermeulen M. 2016. Characterizing protein–protein interactions using mass spectrometry: challenges and opportunities.

Trends Biotechnol. 34(10):825–834. doi:10.1016/j.tibtech.2016.02.014.

Sugishima M, Okamoto Y, Noguchi M, Kohchi T, Tamiaki H, Fukuyama K. 2010. Crystal structures of the substrate-bound forms of red chlorophyll catabolite reductase: implications for site-specific and stereospecific reaction. J Mol Biol. 402(5):879–891. doi:10.1016/j.jmb.2010.08.021.

Sun Y, Shang L, Zhu Q-H, Fan L, Guo L. 2022. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 27(4):391–401. doi:10.1016/j.tplants.2021.10.006.

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, *et al.* 2021. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 49(D1):D605–D612. doi:10.1093/nar/gkaa1074.

Tanaka A, Tanaka R. 2006. Chlorophyll metabolism. Curr Opin Plant Biol. 9(3):248–255. doi:10.1016/j.pbi.2006.03.011.

Thatcher S, Jung M, Panangipalli G, Fengler K, Sanyal A, Li B, Llaca V, Habben J. 2023. The NLRomes of *Zea mays* NAM founder lines and *Zea luxurians* display presence–absence variation, integrated domain diversity, and mobility. Mol Plant Pathol. 24(7):742–757. doi:10.1111/mpp.13319.

Törönen P, Medlar A, Holm L. 2018. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 46(W1):W84–W88. doi:10.1093/nar/gky350.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al.* 2020. Scipy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 17(3):261–272. doi:10.1038/s41592-019-0686-2.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 101(1):5–22. doi:10.1016/j.ajhg.2017.06.005.

Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES. 2014. Association mapping across numerous traits reveals patterns of functional variation in maize. PLoS Genet. 10(12):e1004845. doi:10.1371/journal.pgen.1004845.

Wang S, Wang Y. 2022. Harnessing hormone gibberellin knowledge for plant height regulation. Plant Cell Rep. 41(10):1945–1953. doi:10.1007/s00299-022-02904-8.

Wang S, Wu R, Lu J, Jiang Y, Huang T, Cai Y-D. 2022. Protein–protein interaction networks as miners of biological discovery. Proteomics. 22(15–16):e2100190. doi:10.1002/pmic.202100190.

Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. 2017. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell. 29(5):944–959. doi:10.1105/tpc.17.00009.

Woodhouse MR, Cannon EK, Portwood JL, Harper LC, Gardiner JM, Schaeffer ML, Andorf CM. 2021. A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biol. 21(1):385. doi:10.1186/s12870-021-03173-5.

Woodhouse MR, Sen S, Schott D, Portwood JL, Freeling M, Walley JW, Andorf CM, Schnable JC. 2021. Qteller: a tool for comparative multi-genomic gene expression analysis. Bioinformatics. 38(1):236–242. doi:10.1093/bioinformatics/btab604.

Yu C-W, Liu X, Luo M, Chen C, Lin X, Tian G, Lu Q, Cui Y, Wu K. 2011. HISTONE DEACETYLASE6 interacts with FLOWERING LOCUS D and regulates flowering in Arabidopsis. Plant Physiol. 156(1):173–184. doi:10.1104/pp.111.174417.

Yun J, Kim Y-S, Jung J-H, Seo PJ, Park C-M. 2012. The AT-hook motif-containing protein AHL22 regulates flowering initiation by modifying FLOWERING LOCUS T chromatin in Arabidopsis. J Biol Chem. 287(19):15307–15316. doi:10.1074/jbc.M111.318477.

Zahn T, Zhu Z, Ritoff N, Krapf J, Junker A, Altmann T, Schmutzer T, Tüting C, Kastritis PL, Babben S, *et al.* 2023. Novel exotic alleles of *EARLY FLOWERING 3* determine plant development in barley. J Exp Bot. 74(12):3630–3650. doi:10.1093/jxb/erad127.

Zhang X, Guan Z, Wang L, Fu J, Zhang Y, Li Z, Ma L, Liu P, Zhang Y, Liu M, *et al.* 2020. Combined GWAS and QTL analysis for dissecting the genetic architecture of kernel test weight in maize. Mol Genet Genomics. 295(2):409–420. doi:10.1007/s00438-019-01631-2.

Zhao X, Liu N, Shang N, Zeng W, Ebert B, Rautengarten C, Zeng Q-Y, Li H, Chen X, Beahan C, *et al.* 2018. Three UDP-xylose transporters participate in xylan biosynthesis by conveying cytosolic UDP-xylose into the Golgi lumen in Arabidopsis. J Exp Bot. 69(5):1125–1134. doi:10.1093/jxb/erx448.

Editor: J. Holland