The Plantae Visualization Platform: a comprehensive web-based tool for the integration, visualization, and analysis of omic data across plant and related species

Antonio Santiago[1]*, Luis Orduña[1], José David Fernández[2], Álvaro Vidal[1,3,4], Iñigo de Martín-Agirre[1], Purificación Lisón[5], Elena A. Vidal[2], David Navarro-Payá[1+], José Tomás Matus[1+]

[1] Institute for Integrative Systems Biology (I2SYSBIO, UV-CSIC), Valencia, Spain

[2] Center for Genomics and Bioinformatics, Universidad Mayor (Chile)

[3] Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

[4] Center Agriculture Food Environment (C3A), University of Trento, San Michele all'Adige, Italy

[5] Instituto de Biología Molecular y Celular de Plantas (IBMCP), Valencia, Spain

[*] **First author**

[+] **Corresponding authors**

Abstract

The increasing availability of omic data from non-model plant species has created a need for centralized, user-friendly platforms that maximize the utility of these datasets. Here, we introduce PlantæViz, a web-based tool designed for the integration, visualization, and analysis of multi-omic data across a wide range of plant species. PlantæViz offers advanced functionalities including gene co-expression and regulatory networks, gene expression atlases, cistrome visualization and metabolomics-transcriptomics integration, providing a robust framework for hypothesis generation and biological interpretation. A suite of GeneCards applications tailored to each plant species provide both community curated and automatic functional annotation information. One of the main features of the platform is its big data approach, whereby large volumes of publicly available transcriptomic data have been processed and displayed, i.e. over 58,000 transcriptomic samples are included in total. In addition, a particular effort in the automatic classification and standardization of omic metadata through regular expressions and data mining has been made, with over 90% of transcriptomic runs being successfully classified according to sample tissue. This data has been used to build gene networks through computationally intensive approaches using different methods. PlantæViz connects genomic and functional understanding between model and non-model plant species and aims to expand its catalog of species in the future. Further developments will include the integration of enhanced data types, and novel tools to further support plant research across diverse biological contexts.

Introduction

The rapid accumulation of omic data in non-model plant species has reached levels comparable to those of well-studied model organisms in the recent past. For instance, Illumina-based transcriptomic data in grapevine currently exceeds 10,000 runs in the NCBI Sequence Read Archive (SRA) (Sayers et al., 2022), a milestone Arabidopsis only achieved in 2017. Such extensive datasets highlight the need for up-to-date, centralized web-based platforms for these less-studied plant species to maximize efficient data integration, reuse, and further analysis. The absence of such platforms prevents researchers from fully benefiting from the collective knowledge generated by the scientific community.

While the volume of omic data in non-model crop species may be comparable to model organisms a few years ago, the level of biological insight remains far behind, largely due to the smaller size of the research communities involved. *In silico* approaches, such as sequence-based functional annotation or gene co-expression analyses, offer effective strategies to bridge this gap by inferring putative gene functions. Gene co-expression networks, based on the 'guilt-by-association' principle, have been widely employed in non-model crops (Childs et al., 2011; Orduña et al., 2022). Genes that are consistently co-expressed across diverse conditions are more likely to share functional relationships, even when their specific biological roles remain uncharacterized. These networks provide researchers with a powerful means to predict the potential functions of poorly annotated genes based on their associations with well-characterized counterparts.

The integration of multi-omic data spanning genomics, transcriptomics, and metabolomics is central to systems biology approaches, offering a comprehensive view of an organism's biology. However, the sheer scale of these datasets often leads to their underutilization, primarily due to the lack of accessible, user-friendly tools for visualization and reanalysis. Omic data holds immense potential for exploring gene expression patterns across various tissues, developmental stages, and environmental stimuli, providing deeper insights into gene regulation and its impact on phenotypic traits that often extend beyond the scope of initial study hypotheses.

Model organisms such as Arabidopsis benefit from robust, well-coordinated databases such as The Arabidopsis Information Resource—TAIR (Huala et al., 2001). TAIR integrates omic resources using standardized gene nomenclatures, genome assemblies and structural and functional annotation files as well as providing diverse tools for exploring large-scale data, including mutant collections, literature data, gene expression behavior, etc. In contrast, existing databases for non-model plants are often limited in scope, lacking advanced tools such as gene co-expression analyses or tissue- and condition-specific exploration that could assist researchers in hypothesis generation and testing. Moreover, species-wide platforms such as STRING (Szklarczyk et al., 2023) often fall short for less-characterized

plants, as they rely on outdated annotations, obsolete gene IDs from older genome assemblies, or microarray probe-specific identifiers.

In this study, we introduce PlantæViz (http://plantaeviz.tomsbiolab.com/), a versatile and centralized platform for visualizing and analyzing transcriptomic and metabolic data across a range of plant species, grouped in data modules (named suites) that include Vitis vinifera, Solanum lycopersicum, Morus alba, and Cannabis sativa, Arabidopsis thaliana on its first release. Unlike model-organism-centric platforms, PlantæViz accommodates up-to-date genome assemblies and annotations for these species, which are often underrepresented in existing tools. The development of PlantæViz has involved extensive manual curation, including literature-based compilation of gene catalogues and the creation of hierarchical ontologies to ensure consistent metadata classification. Notably, the platform integrates Arabidopsis data as a functional gateway, leveraging its extensive gene characterization to facilitate insights into less-studied plant species. New suites are envisioned to appear in future releases, further expanding the platform's capabilities to meet changes in research needs. PlantæViz provides an intuitive, user-friendly interface equipped with tools to generate insightful visualizations such as heatmaps, line plots, and network diagrams aimed at enhancing data interpretation and hypothesis generation. By addressing a critical gap in available tools for non-model plants, this platform allows plant researchers to maximize the utility of omic datasets and advance the understanding of gene function and regulation in economically and agriculturally important crops.

Results

A modular design

PlantæViz (http://plantaeviz.tomsbiolab.com/) is organized into different sub-platforms or suites that can be intuitively accessed from the homepage (Figure 1). The homepages of each suite (VitViz, MulViz, TomViz, CannaViz and AraViz) follow a standardized format which allows both user-friendly navigation and a modular architecture, favoring personalized future updates for additional species. Each homepage provides an informative index including menus that lead to the different apps and tools which encompass gene network visualization, gene expression atlases, differential gene expression, gene set enrichment analysis, genome browsers, multiple-sequence alignment, gene function and genomic information, and updated blast server capabilities.

Figure 1. An overview of the different subplatforms or suites in PlantæViz.

Transcriptomic metadata annotation and processing

The percentage of assigned SRA runs varied across categories (cultivar, stage, tissue, and treatment type) for the five analyzed species (*Vitis spp.* for grapevine, *S. lycopersicum* for tomato, *Morus spp.* for mulberry, *Cannabis spp.* for Cannabis, and *A. thaliana* for thale cress) based on the implemented ontology framework (Supplemental File 1). Transcriptomic run metadata classification results are displayed in three different columns per category: 1) for all terms automatically identified by the pipeline ("_curated_details" columns) 2) for the single-most specific term ("_curated" columns) 3) selected ontology term levels for summary ("simplified_curated" columns) (Supplemental Table 1, sheet 1). The selected ontology terms for summarising had an impact in downstream data generation such as network construction since, for instance, choosing "flower" as a summary term would mean that any runs belonging to "flower" subcomponents defined in the ontology would be effectively treated as the summarised term, i.e. "flower". Moreover, "pollen" was often added as a selected summary term so that transcriptomic runs belonging to this very specific sub-tissue would not be treated as a "flower" run downstream.

Overall, the assignment success was higher in certain categories, particularly the "tissue" category exhibited the highest assignment success, with near-complete annotation coverage across most species (Figure 2A). This strong performance can be attributed to the "tissue" ontology being the most developed and refined to date, enabling more accurate assignment of runs, and highlighting the robustness of the ontology-driven regular expression approach in this category. Further analysis of tissue type assignments revealed substantial variability between species (Figure 2B). In Arabidopsis, tissue categories such as leaf, root, and seedling were the most represented, alongside notable contributions from vascular tissues and cell cultures, reflecting the experimental diversity for this model species. In Cannabis, tissues associated with flowers and trichomes were sampled more frequently, aligning with the known research emphasis on secondary metabolite production. For grapevine, tomato and mulberry, the majority of samples were classified as leaf or fruit tissues, consistent with many studies focused on fruit development and associated metabolic

pathways. However, root tissues were underrepresented in grapevine, likely due to the prevalent use of roots from rootstock species rather than the main species itself.

In contrast, assignment rates for the "cultivar" and "stage" categories showed greater variability (Figure 2A). Arabidopsis, grapevine, and cannabis achieved higher assignment success compared to tomato and mulberry, where rates dropped below 50% in some cases. Specifically, tomato showed lower assignment success in "stage" and "cultivar" categories, possibly reflecting greater heterogeneity or incompleteness in metadata. For the "treatment type" category, currently under development using tomato as a proof of concept, assignment success reached approximately 50%, indicating promising progress despite its early stage.

This pattern underscores the challenge of curating and standardizing unstructured metadata, particularly in non-model species where incomplete or inconsistent metadata remains prevalent. While the ontology-based approach (Figure 2C) effectively improved SRA run reannotation across multiple categories, its performance was contingent on both the completeness and the specific characteristics of each species' datasets.

A total of 32,626, 13,719, 10,906, 657, and 443 transcriptomic samples were downloaded and aligned for *Arabidopsis thaliana*, *Solanum lycopersicum*, *Vitis vinifera*, *Cannabis sativa*, and *Morus alba*, respectively (Supplemental Table 2). Expression data for samples with assigned tissue categories were visualized in GeneCards and GeneXplore applications, providing an overview of the annotated transcriptomic runs. A curated subset of these data was further employed for downstream analyses, including network generation.

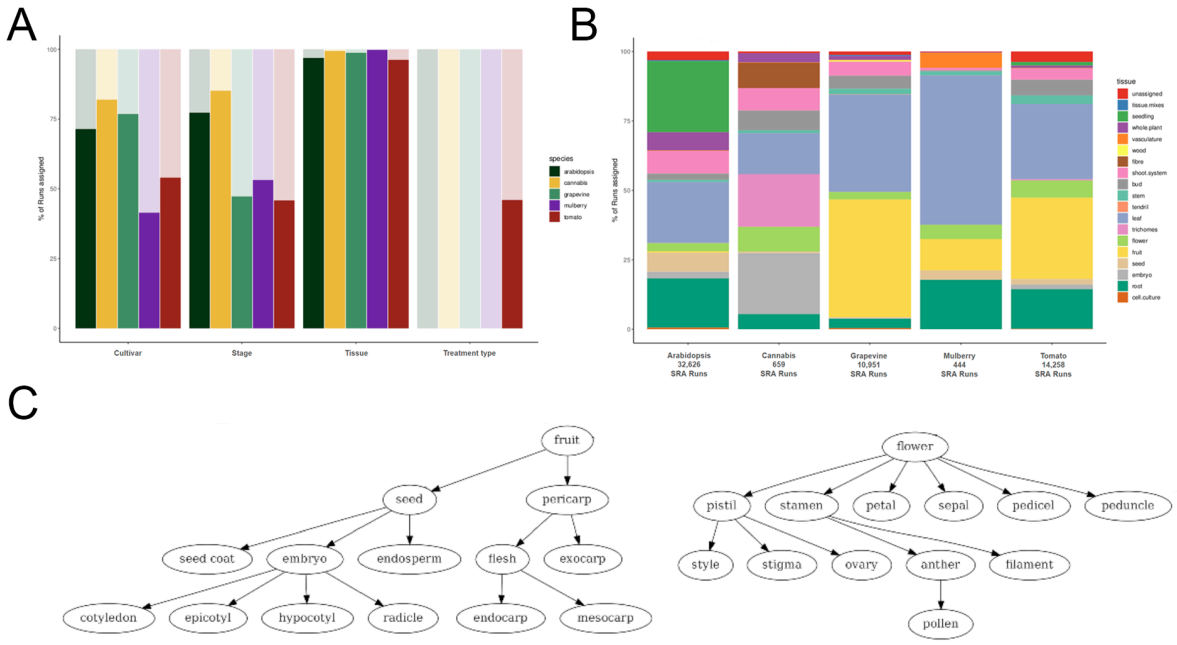

Figure 2. Ontology-driven reannotation of SRA runs across categories and species. (A) Percentage of assigned SRA runs across different metadata categories and species. The tissue category is taken as a model for the rest of categories still in development, exhibiting the highest assignment success across all species. (B) Distribution of tissue type assignments varied substantially between species, showing variability in tissue representation, which appears to match species-specific research interests. (C) A subset of the tissue ontology, illustrating relationships of fruit and flower components. The established hierarchical relationships permit the optional grouping of runs in higher categories even if the sampled tissue pertained to a subcomponent.

GeneCards app

GeneCards applications across all species include automatic functional annotation for queried genes (Table 1), integrated with structural models in JBrowse2, TPM expression values across all metadata-annotated RNA-seq tissue samples from SRA, and the corresponding gene, transcript, CDS, and protein sequences (Figure 3). For grapevine, the GeneCards application is tightly linked to the grapevine gene reference catalogue (Navarro-Payá et al., 2022) (available at https://grapedia.org/genes/), a comprehensive and manually curated resource containing functional data for over 3,000 grapevine genes which is open to submissions through the Grapedia portal.

Table 1. Genome versions and automatically functionally annotated gene counts.

Organism	Genome assembly	Annotation version	Structural genes	Mapman	Interpro	GO	EggNOG Descriptor
Tomato	ITAG4	ITAG4.0-4.2	37467	16795	30577	14099	29677
Arabidopsis	TAIR10	Araport11	38089	15878	25204	15618	23560
Mulberry	Jiao et al. 2020	Mal_HE_154X_1	31401	13547	26333	14340	25468
Grapevine	T2Tv5	v5.1	47972	16016	29678	13069	27411
Cannabis	JamaicanLionDASH; CBDRx-18; PurpleKush	JamaicanLionDASH; CBDRx-18; PurpleKush	61143; 33371; 47954	25199; 13829; 18632	55221; 29147; 42359	24816; 13245; 19651	48539; 25487; 37236

The gene catalogue is built upon the foundation of the telomere-to-telomere (T2T) v5 cv. “PN40024” genome assembly, where V5.1 gene IDs serve as the primary identifiers for genes. Grapevine currently serves as the primary model species for the platform with ongoing efforts to extend similar curation processes to additional species, thereby broadening its scope and utility across diverse plant systems. The GeneCards application acts as a central point of access to the catalogue and retrieving detailed, curated information for individual genes. Users can search by gene symbol or V5.1 ID to access a gene-specific summary sheet, which includes: gene symbols and synonyms, full descriptive gene names, pathway and gene family classifications, and validation levels for functional data.

To support the research community's transition to updated gene annotations, the platform features a cross-annotation tool. This enables researchers to input a single gene or a list of genes and retrieve corresponding identifiers across all available annotation versions, ensuring seamless adaptation and interoperability.

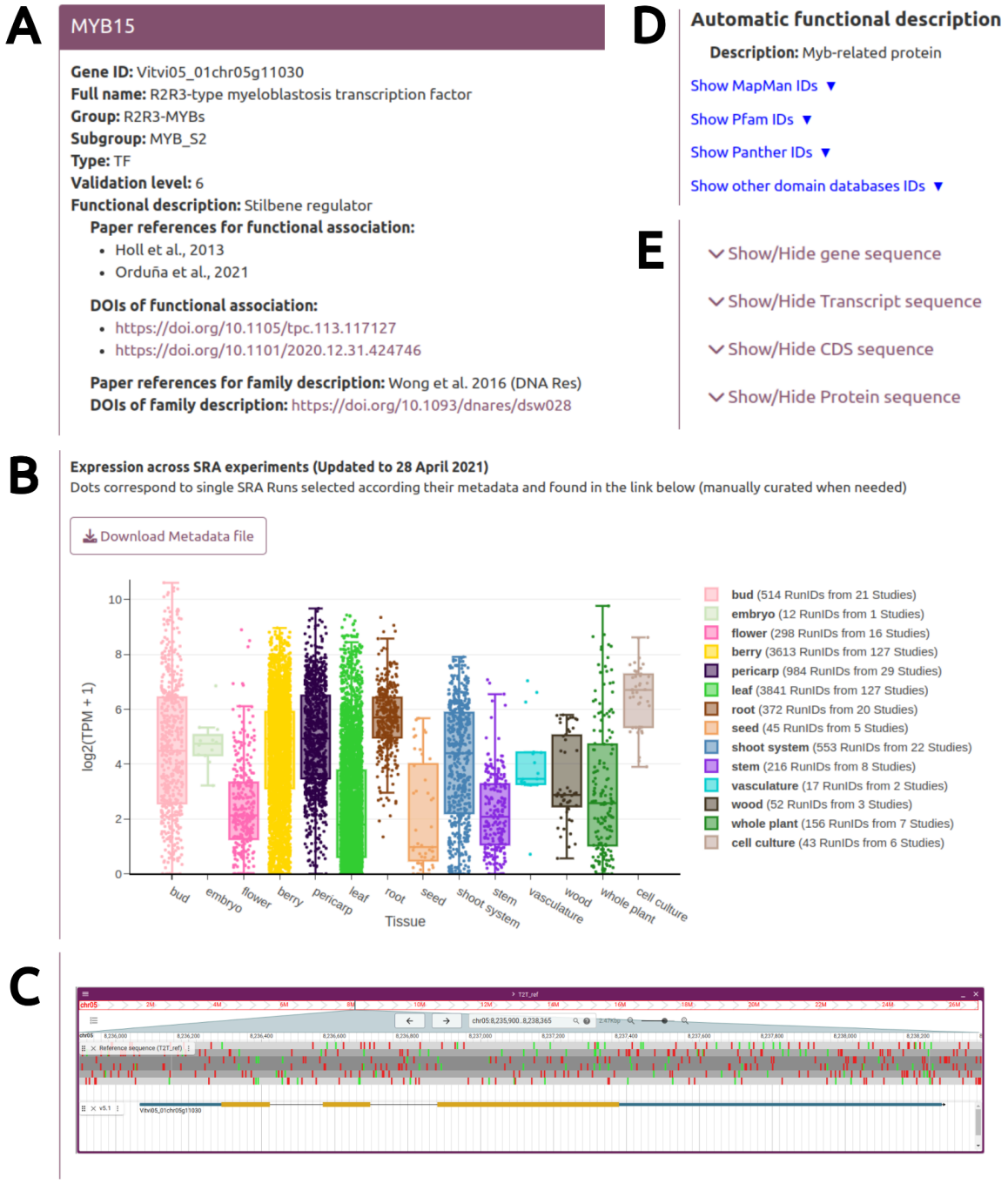

Figure 3. VviMYB15 as an example in VitViz GeneCards. Gene catalogue information (A) is exclusive to VitViz for now, but the rest of the panels (B-E) are available for all species. (A) Manually curated information corresponding to grapevine gene reference catalogue. (B) Tissue-specific gene expression in $\log_2$(TPM+1) for over 10,000 SRA runs, the interactive plot allows the indentification of a particular SRA study by hovering over each point. (C) JBrowse2 windowed instance for the gene of interest. (D) Automatic functional annotation, with category-specific dropdown lists. (E) Fasta format gene, transcript, CDS or protein sequences are also provided.

Gene co-expression and regulatory network apps

We developed an interactive co-expression and regulatory network visualization tool that combines multiple functionalities to explore gene relationships and regulatory interactions. The application is organized into two main tabs, tailored to facilitate different aspects of network analysis while offering advanced visualization features for in-depth exploration. The platform integrates co-expression data with regulatory predictions, including transcription factor (TF) binding events at open chromatin regions, to construct regulatory networks. This dual approach enables users to retrieve TF-target gene pairs, visualize specific subnetworks based on custom gene sets and explore co-expression and regulatory relationships simultaneously.

The app consists of a co-expression table and chromosomal mapping section and another for 3D Network Visualization. In the first tab, users can query a gene of interest to retrieve a table of co-expressed genes based on a predefined co-expression threshold (by default set at 1% TOP). This table provides detailed information such as gene identifiers and names (imported from gene catalogues) and co-expression metrics. Furthermore, users can perform gene set enrichment analysis (GSEA) directly on the list of co-expressed genes to identify enriched biological pathways or processes. Additionally, the chromosomal locations of the co-expressed genes are visualized on a chromosomal map. This feature allows users to examine genomic distribution patterns, in a seamless connection between co-expression analysis and genomic context, highlighting potential chromosomal clustering or co-localization of functionally related genes.

The second tab provides an interactive 3D graph-based visualization of gene co-expression or regulatory networks. A key feature includes custom input lists, where users can input specific gene sets to examine their connections based on co-expression data or regulatory interactions. The network is displayed as a dynamic, three-dimensional graph where genes are represented as nodes and edges reflect co-expression or regulatory relationships. This tool is particularly useful for visualizing pathway or family-oriented networks, revealing how those genes are interconnected and potentially co-regulated. Users can thus identify hub genes (i.e., potential key regulators), and closely connected clusters within their input list. The dynamic graph (3D interactivity) allows users to manipulate views in real time, facilitating deeper insights into the structural organization of the gene network. The topology of the subnetwork is influenced by node connectivity: node positions are a compromise of push/pull forces, they repel each other by default and attract each other based on the number of edges and their weights. Users can adjust repulsion forces to optimize the visualization for their specific network topology.

Expression atlases

The platform incorporates expression atlases applications that consolidate transcriptomic data from various studies, including microarray and RNA-seq datasets. These applications serve as a centralized resource for exploring gene expression patterns across a wide range of biological contexts, such as specific tissues, developmental stages, or experimental conditions. Among its key features, users can create interactive heatmaps from specific gene input lists. Users can also convert expression values to logarithmic or Z-score scales, and perform clustering; group data by genes or conditions to reveal similar expression patterns. Atlases can show normalized expression (TPM) or differential expression data, in the latter, cell color depicts fold-change values while dot size reflects adjusted p-values. These features enable robust comparative analyses, allowing researchers to identify expression trends and potential functional relationships across biological conditions.

ChIP and DAPBrowse

This resource enhances the exploration of TF-DNA interactions across plant species. It consists of genome browsers (JBrowse adaptations) that integrate community-contributed chromatin immunoprecipitation (ChIP)-seq and DNA-affinity purification (DAP)-seq data for transcription factors (TFs), offering a centralized repository for DNA-binding insights. Supplemental Table 3 summarizes the transcription factors currently available in ChIP/DAPBrowse, with the collection expected to expand as additional datasets are contributed by the community.

Annotation tracks are color-coded depicting coding exons displayed as yellow boxes, UTRs as blue boxes, and introns as connecting lines. For each transcription factor, ChIP/DAPBrowse provides three distinct track types: 1) Individual replicate tracks visualizing alignment distributions of reads for TF or input libraries. 2) Merged replicate tracks to compare aggregated read alignments of the TF versus input libraries. 3) GEM peak-caller tracks to highlight peaks specific to a given transcription factor.

Compared to tabular peak data, which can introduce false negatives or positives during automatic peak detection, visualization within the genome browser offers significant advantages by visualizing the alignment tracks within the genome browser, users can manually inspect alignment tracks to confidently validate peaks. Additionally, GEM peak-caller tracks link detected peaks to the most likely associated genes using a default nearest-genomic-feature assignment approach. By combining robust alignment visualization with peak identification tools, ChIPBrowse and DAPBrowse provide an intuitive and reliable framework for studying transcription factor binding dynamics and their regulatory impact.

Multiple sequence alignment

To enhance the exploration of sequence conservation and functional insights across related genes, we developed an application for multiple sequence alignment (MSA). This tool enables researchers to align nucleotide (mRNA or CDS) or protein sequences, identifying conserved regions, motifs, and sequence variations critical for understanding gene function and evolution. The MSA application supports a list of gene IDs for the species of interest, provided by the user. It also offers multiple alignment algorithms, including ClustalW, Clustal Omega, and MUSCLE. These supported methods allow customization of the alignment process to suit specific needs.

The results are displayed in an interactive color-coded alignment viewer, with conserved residues or nucleotides highlighted in a logo graphic to emphasize conservation patterns across aligned sequences. Users can zoom in on specific regions or sequences of interest for detailed inspection. Finally, alignments can be exported in FASTA format for further downstream analyses. By combining robust alignment tools with user-friendly visualization, the MSA application provides a streamlined solution for studying sequence conservation, functional domains, and variations across related genes.

TransMetaDb

TransMetaDb is a tool designed for the integrated visualization and analysis of transcriptomic and metabolomic data, enabling users to explore correlations between gene expression and metabolite levels across diverse datasets. This application provides an accessible platform for uncovering functional associations, linking genes to metabolite traits under specific conditions or experiments. The tool generates a downloadable output in the form of a heatmap or table. In this output, rows represent the input genes, while columns correspond to different metabolites. Each cell displays the Pearson's correlation coefficient (PCC) and its p-value (in brackets), calculated using the Student asymptotic method. Additionally, the application includes a second feature for exploring co-expression clustered modules obtained via a weighted gene co-expression network analysis (WGCNA) performed locally. This functionality enables users to download a table categorizing all genes into their respective modules, offering deeper insights into gene-metabolite associations.

GeneXplore

To facilitate the exploration of the normalized expression data derived from SRA samples, we developed a user-friendly web application that allows users to visualize and compare expression levels of genes of interest across various experimental categories. This tool integrates the ontology-based annotations described above and displays normalized gene expression values (transcripts per kilobase of exon model per million mapped reads, TPM) to improve comparability across samples. The application enables users to examine expression differences for specific genes

across tissues, developmental stages, or treatment conditions, in a combinatorial manner depending on the curated and available metadata. This functionality provides a powerful resource for identifying patterns and gaining insights into gene behavior under diverse biological contexts.

Materials and methods

High performance computing

A high-performance computing (HPC) cluster, Garnatxa (I2SysBio), was necessary to allow for transcriptomic data processing and gene network generation in a timely fashion, given the big volumes of data involved. The cluster consists of 464 CPU cores, 14 Tb of RAM memory, of which a maximum of 200 CPUs, and 1.2 Tb of RAM were used at any one time. Parallelization of Sequence Read Archive (SRA) run downloads made use of the system's SLURM manager.

Download data and metadata classification

Retrieval of raw RNA-Seq data from SRA was focused explicitly on Illumina sequencing runs, downloaded using the SRA Toolkit, filtering only for transcriptomic data experiments and excluding datasets related to micro, small, or non-coding RNAs. To identify relevant RNA-Seq datasets, and fulfil these conditions the following species-specific queries were used, see Supplemental Table 4. Wherever necessary, just the genera of the species was used to capture data from very closely related accessions. The Arabidopsis query has an additional filter to include only paired-end studies given the volumes of data available.

Retrieved metadata was automatically classified using a species-specific ontology. This ontology, formatted as a JSON file, consists of hierarchical terms organized into the following categories: tissue, organism (species), stage, cultivar and treatment. Each term is associated with a set of search terms and regular expressions used to identify its presence in the metadata. Only relevant metadata columns were considered for each category and species (Supplemental Table 1, sheet 2). Classified metadata terms were subsequently summarized into more general upper-level terms from the ontology, see Supplemental Table 1 sheet 1 for selected terms, to facilitate downstream analysis and processing.

GeneCards

Functional annotation was generated for each species using both eggNOG-mapper and Mercator4, and then reformatted to GMT format for gprofiler2 GSEA analysis, the input format for other apps. In the case of grapevine, curated gene functional information was integrated based on the submission forms provided by the research community (https://grapedia.org/genes/). Gene ID correspondences between grapevine annotations were established by using the Annotation extraction and gene integration suite (AEGIS), which was developed *in-house* and is based on a

weighted multi-feature coordinate overlap algorithm. A JBrowse2 genome browser was set up for each species considering the latest genome assembly and annotation versions (Table 1). An integrated JBrowse2 window was included using the inbuilt language for gene coordinate queries.

Gene network construction

The network construction involved a series of steps to collect, process, and analyze RNA-Seq data obtained from the NCBI SRA database, which was independent of the type of network built; non-aggregated gene co-expression network (non-agg GCN), aggregated GCN (agg GCN), or gene regulatory networks (GRNs). All RNA-Seq reads were processed in a series of steps to prepare them for alignment to the reference genome. This involved trimming the reads using the fastp (version 0.20.0) (default parameters + *--cut_front_window_size 1 --cut_front_mean_quality 30 --cut_front cut_tail_window_size 1 cut_tail_mean_quality 30 --cut_tail -l 20*) to remove low-quality sequences and adapter contamination. Subsequently, the trimmed reads were aligned against the corresponding genome assembly for each species (Table 1) using the STAR aligner (version 2.7.3a) (default parameters + *--runMode alignReads --limitOutSJcollapsed 8000000 --limitIObufferSize 220000000*).

Following alignment, raw count matrices were generated for each SRA run using featureCounts (version 2.0.0) and the respective annotation version for the species (Table 1). To account for variations in sequencing depth across different runs, the raw count matrices were normalized using the transcripts per kilobase of exon model per million mapped reads (TPM) method. Genes exhibiting consistently low expression levels, defined as less than 0.5 TPM across all runs within a study, were removed from further analysis.

The process of constructing aggGCNs and non-aggGCNs involved a combination of correlation analysis and network aggregation techniques as described in (Orduña et al., 2023) with some modifications in the pipeline that permit SRA study tracking for aggGCNs. In the case of aggGCNs, the measure of co-expression is based on the number of SRA studies where a co-expression relationship can be ascertained for any pair of genes. Knowing the nature of these studies and not only the total number can be particularly useful to uncover potential condition-specific relationships. The construction of inferred GRNs, were based on the GENIE3 algorithm implemented in the GENIE3 R package (Huynh-Thu et al., 2010). Reference GRNs (rGRNs) included additional *in silico* and experimental sources of evidence to qualify regulatory relationships into 4 confidence levels. The D3 network R package was modified and implemented within the different network apps to show interactive 3D plots based on user-defined gene lists, allowing for focused exploration of co-expression relationships within a specific set of genes. A custom slider to modify repulsion forces is provided. The chromosome map utility was custom built with ggplot2. D3 subnetwork visualization for GRNs was modified to allow for directional

edges in the form of arrows (suggestive of regulation) and a color-coded system was implemented for rGRNs to indicate the different levels of evidence supporting a particular relationship.

DAP-Browse

Sequencing reads obtained from DAP-seq and ChIP-seq experiments were aligned to the corresponding species reference genome using Bowtie2 (Langmead & Salzberg, 2012) with default settings. Reads with MAPQ scores below 30 were filtered out during post-processing to ensure data quality. Peak detection was carried out with the GEM peak caller (Guo et al., 2012) using the respective genome assembly and specifying the following parameters: -q 1 -t 1 -k_min 6 -kmax 20 -k_seqs 600 -k_neg_dinu_shuffle. Analysis was limited to nuclear chromosomes, and replicates were processed using GEM's multi-replicate mode. Peak summits identified by GEM were mapped to the nearest gene models from a custom annotation file using the BioConductor package ChIPpeakAnno (Zhu et al., 2010) under default parameters (NearestLocation).

To facilitate downstream analysis and visualization, the detected peak coordinates were converted to a BED file format. Alongside the alignment-derived bigWig file, these data were uploaded into the JBrowse2 genome browser for integration and inspection.

Differential expression analysis and metabolite integration

Transcriptomic and metabolomic integration was performed as described for the TransMetaDB app (Savoi et al., 2022). Briefly, expression data was normalized to TPM. Differential expression analysis (DEA) was performed using the LIMMA R package. Genes were classified as differentially expressed when the adjusted p-value (Benjamini-Hochberg correction) was below 0.05. Co-expression network analysis was performed using the Weighted Gene Correlation Network Analysis (WGCNA) R package for metabolite integration.

BLAST servers

Each species' datasets were individually formatted into BLAST-compatible databases using the makeblastdb utility. We implemented a local installation of SequenceServer (v3.1.3), an intuitive and user-friendly platform for running BLAST (basic local alignment search tool) queries on transcriptomic, CDS, and protein data. The user is able to select parameters based on the BLAST command line tool. Output formats include alignments, detailed hit descriptions, and graphical overviews of query matches.

Implementation

PlantæViz was developed as a Shiny server (v1.5.17.973) running R version 3.6.3. The index page is based on React as the JavaScript library. The platform is freely available for use on the web and can be found at http://plantaeviz.tomsbiolab.com/.

Discussion

PlantæViz sets itself apart from existing omic databases through its unique integration of diverse functionalities, improved scalability, and user-focused design. Unlike many public repositories, PlantæViz combines transcriptomic, genomic, and metabolic datasets into a centralized platform, facilitating seamless multi-omics exploration. The platform significantly enhances webpage responsiveness and data handling capacity, ensuring that researchers can efficiently access, analyze, and interpret large-scale datasets. The extensive curation of these datasets, paired with the incorporation of user-friendly tools, allows researchers to perform species-specific or comparative analyses across organisms. For instance, applications such as *GeneCards* and *Expression Atlases* provide detailed insights into gene expression dynamics, while tools like *TransMetaDb* and gene network visualizations offer deeper connections between genes, metabolites, and transcriptional regulation.

A key distinguishing feature of PlantæViz is its comprehensive transcriptomic data coverage across diverse plant species, encompassing in its first release a total of 58,351 accessions from the NCBI SRA database, which has been instrumental in enabling advanced tools like expression and regulatory network visualizations. This core objective has been made possible through the use of high-performance computational resources, ensuring that the platform remains efficient and scalable even with increasingly complex datasets. The processed transcriptomic data has been used to build tissue-specific networks for all species involved and high-performing aggregated gene co-expression networks were built wherever sufficient data was available. Network aggregation has been shown to improve network performance, but it does require larger datasets and therefore greater computational resources (Wong, 2020).

Another critical advantage of PlantæViz lies in its modular and species-agnostic design. This versatility not only ensures compatibility with a wide range of plant species but also enables the incorporation of additional organisms and advanced computational methods as they emerge. By supporting both established and non-model plants, PlantæViz promotes interoperability between different knowledge bases where data accessibility and analytical tools are improved for non-model species whilst retaining key links to well-studied organisms.

The ability to integrate multi-omic datasets, including transcriptomics and metabolomics, underscores the platform's utility for hypothesis generation and

functional genomics. The suite of tools provided enhance the discovery of functional relationships, conserved gene motifs, and gene-metabolite correlations, offering researchers a systems-level understanding of plant biology. This is particularly beneficial for advancing research in plant development, specialized metabolism, and stress responses, areas where non-model species are gaining increasing attention.

Looking ahead, the continuous incorporation of updated datasets, new species, and state-of-the-art analytical methods will further strengthen PlantæViz's role as a vital resource for plant research and breeding innovations. By integrating advanced multi-omics technologies and expanding species representation, the platform will remain at the forefront of systems biology, fostering new discoveries and accelerating translational applications in agriculture and biotechnology.

Data Availability

In compliance with FAIR (findable, accessible, interoperable, and reusable) guidelines, data has been made accessible through the applications themselves, both in raw form and in the various processed subsets that can be obtained from the various online tools. Moreover, other data and metadata used to create the platform, such as the ontology framework, are provided as supplementary files.

Acknowledgments

The generation of the networks and all the bioinformatic analyses were performed on the HPC cluster Garnatxa at the Institute for Integrative Systems Biology (I2SysBio, UV-CSIC, Spain).

Funding Statement

This work was supported by the grant Valinet (PID2021-128865NB-I00) awarded to J.T.M from the Ministerio de Ciencia, Innovación y Universidades (MCIU, Spain), Agencia Estatal de Investigación (AEI, Spain), and Fondo Europeo de Desarrollo Regional (FEDER, European Union), and to the PROMETEO grant (PROMETEO/2021/056) awarded to P.L. and J.T.M. and its associated doctoral grant (PROMETEO/2021/056-01) granted to AS, by the Generalitat Valenciana (GVA).

Bibliography

Childs, K. L., Davidson, R. M., & Buell, C. R. (2011). Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. *PLoS ONE*, *6*(7), e22196. https://doi.org/10.1371/journal.pone.0022196

Guo, Y., Mahony, S., & Gifford, D. K. (2012). High Resolution Genome Wide Binding

Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. *PLoS Computational Biology*, *8*(8), e1002638. https://doi.org/10.1371/journal.pcbi.1002638

Huala, E., Dickerman, A. D., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B. W., Beavis, W., Meinke, D. W., Town, C. D., Somerville, C., & Yon Rhee, S. (2001). The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Research*, *29*(1), 102–105. https://doi.org/10.1093/nar/29.1.102

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, *5*(9), Article 9. https://doi.org/10.1371/journal.pone.0012776

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Navarro-Payá, D., Santiago, A., Orduña, L., Zhang, C., Amato, A., D'Inca, E., Fattorini, C., Pezzotti, M., Tornielli, G. B., Zenoni, S., Rustenholz, C., & Matus, J. T. (2022). The Grape Gene Reference Catalogue as a Standard Resource for Gene Selection and Genetic Improvement. *Frontiers in Plant Science*, *12*, 803977. https://doi.org/10.3389/fpls.2021.803977

Orduña, L., Li, M., Navarro-Payá, D., Zhang, C., Santiago, A., Romero, P., Ramšak, Ž., Magon, G., Höll, J., Merz, P., Gruden, K., Vannozzi, A., Cantu, D., Bogs, J., Wong, D. C. J., Huang, S. C., & Matus, J. T. (2022). Direct regulation of shikimate, early phenylpropanoid, and stilbenoid pathways by Subgroup 2 R2R3-MYBs in grapevine. *The Plant Journal*, *110*(2), Article 2.

https://doi.org/10.1111/tpj.15686

Orduña, L., Santiago, A., Navarro-Payá, D., Zhang, C., Wong, D. C. J., & Matus, J. T. (2023). Aggregated gene co-expression networks predict transcription factor regulatory landscapes in grapevine. *Journal of Experimental Botany*, *74*(21), Article 21. https://doi.org/10.1093/jxb/erad344

Savoi, S., Santiago, A., Orduña, L., & Matus, J. T. (2022). Transcriptomic and metabolomic integration as a resource in grapevine to study fruit metabolite quality traits. *Frontiers in Plant Science*, *13*, 937927. https://doi.org/10.3389/fpls.2022.937927

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000

Wong, D. C. J. (2020). Network aggregation improves gene function prediction of grapevine gene co-expression networks. *Plant Molecular Biology*, *103*(4–5), 425–441. https://doi.org/10.1007/s11103-020-01001-2

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green,

M. R. (2010). ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, *11*(1), 237. https://doi.org/10.1186/1471-2105-11-237