Charting plant gene functions in the multi-omics and single-cell era

Thomas Depuydt[1,2], Bert De Rybel[1,2], Klaas Vandepoele[1,2,3,*]

[1]Ghent University, Department of Plant Biotechnology and Bioinformatics, Ghent, Belgium

[2]Vlaams Instituut voor Biotechnologie, Center for Plant Systems Biology, Ghent, Belgium

[3]Ghent University, Bioinformatics Institute Ghent, Ghent, Belgium

*Correspondence: klaas.vandepoele@psb.vib-ugent.be (K. Vandepoele)

Key words: multi-omics, single-cell, gene function, networks, data integration, machine learning

Abstract

Despite the increased access to high-quality plant genome sequences, the set of genes with a known function remains far from complete. With the advent of novel bulk and single-cell omics profiling methods, we are entering a new era where advanced and highly integrative functional annotation strategies are being developed to elucidate the functions of all plant genes. Here, we review different multi-omics approaches to improve functional and regulatory gene characterization and highlight the power of machine learning and network biology to fully exploit the complementary information embedded in different omics layers. Finally, we discuss the potential of emerging single-cell methods and algorithms to further increase the resolution allowing to generate functional insights about plant biology.

Highlights

- Different omics profiling methods are becoming available for a variety of plant species, offering new opportunities to explore new gene functions at various molecular levels
- The combination of multiple omics data types enables the characterization of different levels of gene regulation for a biological process of interest
- Studies targeting a specific biological process, as well as untargeted multi-omics approaches shed light on novel gene functions in plant biology
- Different integration strategies offer a practical solution to exploit the information captured by complementary omics layers and allow the generation of new testable hypotheses for various biological processes and pathways
- New computational methods, modeling co-expression and regulatory networks, are being developed to accommodate the technological advances in omics profiling and fully exploit the information available in functional genomics databases

Glossary

Classifier: a class of supervised machine learning algorithms designed to assign objects to separate classes based on a set of input features. For example, genes can be classified to either non-responsive, up-regulated or down-regulated classes for a certain growth condition based on a set of genomic properties. Examples of classifiers are random forest models, support vector machines and artificial neural networks.

Co-expression network: a network where genes are represented by nodes and edges represent (or quantify the level of) similarity between two genes' expression profiles. Within the network, **gene neighborhoods** or clusters can be functionally characterized by applying the **guilt-by-association**

paradigm, which states that genes situated closely together in the network are likely to be functionally related.

Differential expression (DE): the difference in relative abundance of macromolecules, such as transcripts or proteins, between tissues, treatments or developmental stages.

Evidence code: specifies the type of evidence supporting a gene's functional annotation. An evidence code is explicit to any ontology-based annotation and covers different types of experimental, curated or computational support.

Gene Ontology (GO): an ontology, structured as a directed acyclic graph, that describes three aspects of gene function: Biological Process (what role does the gene product fulfill in the biological system), Molecular Function (how does it executes that role), and Cellular Component (where in the cell does it function).

Gene regulatory network (GRN): a collection of regulatory interactions (represented as directed edges in a network) between transcription factors and their target genes (both represented as nodes), describing how gene expression is regulated in specific organs and conditions. GRNs can be mapped experimentally using protein-DNA binding essays, or computationally using for example expression-based GRN inference algorithms.

Gene set enrichment analysis: a statistical technique that identifies significantly overrepresented features (e.g. GO terms or transcription factor DNA motifs) present within a predefined gene set (e.g. a set of upregulated genes).

Module: group of genes showing a similar response in a specific profiling experiment that are frequently functionally related or operate in the same pathway.

Multi-omics: The combined analysis of different high-throughput experimental techniques, each profiling a specific layer of molecular information.

Ontology: a controlled vocabulary describing a set of predefined terms and their relationships.

Plant Ontology (PO): an ontology that describes plant anatomy, morphology and different stages of plant development, thereby linking gene annotations to spatiotemporal expression and observed phenotypes.

Plant Trait Ontology (TO): an ontology that describes plant traits at any scale, encompassing nine broad categories.

Precision: a measure describing the quality of a prediction, reporting the fraction of correct predictions to the total number of predictions.

Spatial transcriptomics: a single-cell RNA profiling technology that does not require tissue dissociation and thus preserves spatial information at a cellular or subcellular level.

The onset of large-scale hypothesis generation for plant gene functions

Unraveling gene functions is pivotal to understanding the signaling cascades and pathways that control growth, development and stress responses. In plant biology, the elucidation of gene functions has historically been realized through forward genetics [1]. While tried and proven, forward genetics is a slow and tedious process, as illustrated by the void in the functionally annotated gene space in plants [2]. With the advent of next-generation sequencing in the 2000s, a wide array of new molecular profiling methods arose, paving the way for large-scale hypotheses generation and reverse genetics validation [2,3]. As such arose the belief that a full understanding of all *Arabidopsis thaliana* (arabidopsis) genes would be realized within one to two decades (Arabidopsis 2010 project) [4,5]. Yet, during the last decade it has become clear that such understanding will not be achieved anytime soon. Multiple reasons lay at the basis of this delay, such as the genetic redundancy between loci diminishing the success rate of single-gene perturbation experiments [6,7]. Furthermore, proteins function at different molecular levels and characterizing, for example, transcriptional, translational and post-translational regulation, is non-trivial. Finally, the emergence of new gene types, such as long noncoding RNAs, reveal another layer of regulatory complexity that is largely uncharted [8–10]. Nevertheless, the plant community continues its quest towards the functional elucidation of all genes, where different omics data types and biological networks now play a central role [2,11]. To aid this aim, various **ontologies** (see Glossary) formally describing gene functions have been proposed and applied in plant research (Box 1), which alleviate many issues regarding ambiguous descriptions and assist in querying databases and performing data mining.

Here, we discuss key concepts revolving around the functional characterization of plant genes using a systems biology approach and exploiting **multi-omics** and single-cell data. Furthermore, we raise outstanding questions that remain to be addressed to ensure optimal extraction of biological information encoded within various omics layers.

Multi-omics approaches for functional characterization

Omics technologies can be defined as high-throughput experimental techniques profiling a specific layer of molecular information (Table1). The entities studied can range from genes (genomics) to metabolites (metabolomics) to interactions between molecules (interactomics). Increasingly, researchers aim to include multiple omics data types, which enable them to capture different aspects or obtain increased cellular resolution of the biological process under investigation (Figure 1, Key Figure).

For example, while **differential expression** (DE, Table 1) analysis is a well-established and straightforward technique to study the transcriptional response to various growth conditions, perturbations, or developmental stages, the discrepancy between transcript and protein (differential) expression has been described on multiple occasions (reviewed in [12]). Thus, adding a proteomics layer can provide a deeper understanding of the molecular response under investigation. Similarly, genome-wide transcription factor (TF) binding events can be experimentally profiled using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) or DNA affinity purification sequencing (DAP-Seq) [13]. However, a binding event does not necessarily imply transcriptional regulation [14], and additional evidence for gene regulation can be obtained from DE of putative target genes upon TF perturbation [15–18]. Likewise, genetic mapping approaches such as genome-wide association studies (GWASs) and quantitative trait loci (QTL) linkage mapping studies are powerful statistical frameworks (Table 1), aimed at linking traits to genomic loci and possible causal alleles. However, the associated loci can stretch over large genomic regions of >1 Mb [19] and additional data is required to further prioritize candidate genes [20]. For example, Schaefer and colleagues used functional information derived from **co-expression networks** (Table 1) to prioritize candidate genes in ion accumulation associated QTLs from maize [21].

While the above examples often apply a simple filtering strategy of molecular findings specific to each data type (e.g. comparison of transcriptome-derived and proteome-derived DE gene sets),

more advanced data integration protocols have been developed, yielding new insights about plant functional genomics. To distinguish between different integration strategies, several semantic classifications have been proposed [22–24]. To accommodate the literature discussed below, we have summarized the most important integration strategies applied by the cited literature (Box 2).

Targeted modeling of different omics layers exposes specific regulatory and functional interactions

Recent studies have demonstrated how the integration of complementary omics data types enlarges our understanding of the wiring of biological networks, in a targeted or context-specific setting centered around a specific developmental process or environmental response [25–30]. For example, Zander and co-workers delineated the jasmonate (JA) signaling network in arabidopsis through dynamic profiling of regulatory interactions, chromatin state, transcriptome and (phospho)proteome profiling [31]. They identified direct regulatory targets of MYC2, a JA-response master regulator, alongside its genetic interaction partner MYC3, through a combination of ChIP-Seq and DE analysis. Moreover, using *in vivo* and *in vitro* binding assays additional downstream and/or known JA-signaling TFs were subsequently mapped, resulting in a highly interconnected JA-response **gene regulatory network (GRN).** Functional regulatory **modules** were identified through **gene set enrichment analysis**, revealing distinct functional groups of JA-signaling TFs, including known and novel gene annotations.

Next, transcriptome, proteome and phosphoproteome profiles were captured along a JA-treatment time course experiment. In concordance with other studies, all three modalities identified complementary DE gene sets. However, the authors surprisingly reported that the responsive (phospho)proteins where all, apart from one gene, not currently associated with JA-signaling. This finding is perpendicular to previous findings in maize, where, although the actual associated genes were different between protein and transcript co-expression clusters, the overall degree of MapMan-based pathway enrichment was very similar for most pathways [32]. Likewise, another

study found enriched brassinosteroid (BR) related GO terms among DE transcripts, proteins and phosphoproteins over a BR-treatment time course in arabidopsis [33]. Thus, these combined findings suggest a lack of known JA-associations among JA-responsive proteins, in agreement with the authors' hypothesis [31]. However, further investigation is needed to fully determine whether this discrepancy is species- or pathway-specific, or a combination thereof.

To complement the experimental GRN and DE analysis, Zander and colleagues aimed to model the JA-regulatory landscape through delineation of a transcriptome- and (phospho)proteome-wide GRN using Regression Tree Pipeline for Spatial, Temporal, and Replicate data (RTP-STAR [34]), which allows to build dynamic networks from steady-state data. RTP-STAR is built on top of the popular GENIE3 framework [35], and includes an initial clustering step and inference of regulation type (positive versus negative). Individual networks were learned using each omics layer (i.e. modality-specific networks), from which a union network was derived as the final GRN. Notably, the union GRN encompassed a 28% increase in gene count compared to the transcriptome-derived network, highlighting the complementary information encoded by different molecular levels of the biological system. Validation of the JA-signaling GRN using the experimentally profiled binding events revealed highly varying levels of **precision**, an indication that some parts of the network are more reliable than others. Additionally, T-DNA loss-of-function (LOF) lines for seven (out of 100 tested) MYC2/3 targets, previously unassociated with JA-signaling, exposed a JA root-growth phenotype. Low phenotype frequency is likely due false positive edges, as well as genetic redundancy, obscuring altered phenotypes in LOF mutants.

In another study, Clark and co-workers adopted a multi-omics approach, integrating transcriptome, proteome and phosphoproteome profiling, to explore the dynamic BR molecular signaling network in arabidopsis [33]. The authors developed a novel GRN inference protocol called Spatiotemporal Clustering and Inference of Omics Networks (SC-ION), which builds on RTP-STAR [34] by incorporating different clustering algorithms for temporal versus no-temporal data, and allows

integration of any number of different types of expression profiles prior to network inference, as opposed to merging modality-specific networks [31]. Using SC-ION, an “abundance GRN” was modeled from protein (TF) and transcript (TF/target) expression levels, and a “phosphosite GRN” was modeled from phosphosite (TF) and transcript (target) expression levels. Both networks were modeled across a treatment time axis, allowing prediction of time-specific regulatory events, which exposed multiple clusters specific to the early and late response, as well as feedforward and feedback regulation between time points. In addition, using differential phosphorylation levels in kinase activation domains, a co-phosphorylation kinase signaling network was delineated, and merged with the two SC-ION networks into a single GRN, modeling the temporal cascade of BR response across these different omics levels. This combined GRN showed that the BR response begins with kinase signaling, resulting in modulation of TF phosphorylation and/or abundance, which in its turn activates the transcriptional response. Next, regulatory motifs within the union GRN, such as feedback and feedforward loops, were identified and used for node (gene) prioritization.

Following computational prediction, Clark and colleagues experimentally validated the regulatory and functional roles of several top regulators. For example, using an *in vitro* kinase assay and transient expression in *Nicotiana benthamiana*, a predicted kinase signaling event where BIN1, a known BR-responsive kinase, phosphorylates BES1, was validated. Moreover, ectopic expression of kinase signaling loss-of-function BES1 mutant showed an hypersensitive BR signaling phenotype, validating the upstream edges of BES1 in the network. Additionally, the downstream regulatory edges of BES1 were validated using publicly available experimental protein-DNA binding data, reporting a precision of 61%. By scoring the BR signaling phenotypes of LOF mutants for three additional top-scoring regulators, two were found to impair BR signaling (ANL2 and BRON). Further experimental characterization of *bron* mutants exposed that BRON is a likely negative regulator of cell division in the root meristem in response to BR. Through measuring *BRON* expression in LOF

mutants of six predicted upstream regulators, the predicted *BRON* transcriptional cascade, responsible for BR-induced cell division in root meristem, was experimentally validated.

Next to phytohormone signaling, biotic or abiotic stresses can be investigated using a multi-omics approach, resulting in identification of novel regulators or pathway genes. For example, in a comparative study, researchers investigated the response to excess boron in both arabidopsis and its boron-insensitive close relative *Schrenkiella parvula* [36]. Using comparative genomics and transcriptome profiling under normal and excessive boron conditions, they showed that DE orthologs in arabidopsis are constitutively expressed in *S. parvula*, concluding its transcriptome is pre-adapted to boron toxicity. For example, the DE putative boron transporter *AtBOR5* was not DE in *S. parvula*; however its expression showed a >2000-fold higher constitutive expression. Ectopic expression of *SpBOR5* in a boron transporter deficient yeast line fully rescued growth under excess boron conditions, confirming its role as a boron transporter. Moreover, using pathway-mapping of DE genes and conceptual integration with metabolomic profiling, the authors exposed boron induced alterations to cell wall metabolism in arabidopsis.

In another study, the effect of flooding on the gene regulatory circuitry was investigated and compared across four angiosperm species [37]. By combining chromatin accessibility, gene expression, putative TF-binding information and conservation of genomic context (synteny), a set of evolutionary conserved genes, spanning 68 gene families, were identified as activated by submergence and under coordinated regulation by four TF families. Aside from genes known to be essential in anaerobic metabolism and hypoxia response, such as the *PCO* family, genes not well associated with submergence, such as the *PYR/PYL* family, were identified and might pose novel targets to enhance flooding tolerance in susceptible crops.

To conclude, these studies showed that through a targeted integrative multi-omics approach, hormonal signaling and stress response networks can be delineated and used to predict novel candidate regulators or response genes. In these studies, transcriptomics plays a central role in

exploring the omnipresent transcriptional regulation, and is often applied in conjunction with proteomics, which exposes additional levels of post-transcriptional regulation, such as protein degradation or kinase phosphorylation. Moreover, the different omics layers are combined into GRNs modeling the process under study, which achieve higher levels of accuracy when compared to single-omics networks. For this task, new computational algorithms are actively being developed, allowing the field to progress and keep up with the technological advances in omics profiling. Functional and regulatory hypotheses inferred from these networks are validated using experimental reverse genetics techniques, providing novel high-confidence characterizations for key molecular players in the processes under investigation.

Untargeted approaches shed light on the various functions of unknown plant genes

The studies discussed above are targeted on a single biological process, allowing the researchers to investigate the process under study in a context-specific manner, yielding high-resolution information. Conversely, untargeted approaches have the potential to cover a broad array of processes, though with potentially lower resolution, often making use of publicly available datasets. One strategy to add context-specificity to an untargeted approach is to sample many complementary datasets, each profiling a specific tissue, treatment or developmental stage, and adopt an integration scheme that does not obscure dataset-specific interactions.

For example, Zhou and colleagues compiled a large compendium of publicly available transcriptome datasets in maize, to build a collection of 45 co-expression-based GRNs in maize [38]. Comparison of edges between these networks and published direct (based on *in vivo* DNA-binding and DE analysis using TF knockout lines) or indirect targets (only DE analysis) revealed significant enrichment for targets for four out of six and 14 out of 17 TFs, respectively. The authors reported that developmental networks tend to model many TFs, while tissue-specific networks are superior in predicting TFs that are specific to the corresponding tissue. Moreover, inspection of several well-

characterized metabolic pathways and linked regulatory modules showed that distinct networks capture complementary information for certain pathways. Collectively, these findings exemplify the breadth and biological relevance of an untargeted approach, while retaining context-specific information. Next, the authors queried publicly available QTLs where the trait under investigation is the variation in target gene expression, called expression-QTL (eQTL), and investigated the tendency of the eQTL hotspot associated target genes to share a common predicted regulator (reviewed in [39]), based on the collection of GRNs. Furthermore, genomic colocalization of predicted regulators and eQTL hotspots identified 68 candidate causal TFs underlying 74 eQTL hotspots. Functional enrichment of the regulatory modules recovered known functions for several of these TFs, as well as suggested novel roles for uncharacterized TFs. For example, the COL11 response modules were enriched for photosynthesis, and the MYC7 response modules were enriched for the JA biosynthesis pathway. However, although arabidopsis orthologs for these TFs are involved in related processes, experimental validation is still needed to fully characterize these maize regulators.

The integrated analysis of a large compendium of transcriptomics and interactomics data was used to infer functional annotations for 5,054 arabidopsis genes lacking a GO Biological Process annotation [40]. Similar to [38], a selection of gene expression datasets were used to build 18 gene co-expression networks, modeling different developmental stages, sampled tissues and experimental treatments. These individual networks were first jointly interrogated using variable-Highest Reciprocal Ranks (vHRR), a novel network propagation protocol which attributes biological processes to the most relevant co-expression networks and optimizes queried co-expression neighborhood sizes on a process-by-process basis. This approach retains condition-specific co-expression relationships while simultaneously maintaining the breadth of a condition- or process-independent approach [41]. Secondly, validation using experimental protein-DNA and protein-protein interaction networks added a physical and/or regulatory context to the co-expression-based functional hypotheses. As such, novel functions for many unknown genes were identified for variety

of developmental processes and molecular responses, such as flower and root development, defense responses to fungi and bacteria, and phytohormone signaling [40]. Moreover, an in-depth analysis of a drought response subnetwork showed that five uncharacterized genes may facilitate crosstalk between abscisic acid and cytokinin signaling in arabidopsis. Similarly, a seed development subnetwork revealed various potential roles for multiple uncharacterized genes during seed maturation.

Taken together, by leveraging a large and diverse collection of public transcriptomics data and applying an appropriate integration method, often on a network level, transient interactions can be retained. Moreover, adding omics layers such as trait-associations or interactome data can provide additional evidence for transcriptome-derived functional hypotheses, resulting in high-confidence candidates for experimental validation and potential application for crop improvement. Indeed, the use of public data reduces the cost and time needed to unravel gene functions, and importantly, it allows to leverage available data sets, effectively uncovering valuable information that was left unexplored. Advances in computational techniques ensure that relatively established approaches, such as co-expression analysis, continue to evolve and incorporate the increasingly diverse and extensive collection of public data sets.

Machine learning models allow for simultaneous integration of omics features and regulatory or functional classification of genes

To fully exploit the complementarity of different molecular profiling methods, machine learning offers a powerful approach to infer gene functions [11,42,43]. Recently, De Clercq and co-workers applied a network-based approach for large-scale integration of different functional data types, with a major goal to enhance our understanding of TF gene regulation in arabidopsis [44]. A supervised learning approach (reviewed in [43]) was used to first train a machine learning algorithm, called a **classifier**, exploiting information about TF DNA motifs, open chromatin, TF-binding and expression-

based regulatory interactions. Subsequently, the classifier was used to prioritize and identify those interactions that most probably represent functional interactions. The regulatory interactions retained by the classifier resulted in an integrated genome-wide GRN (called iGRN), covering 1,491 TFs and 31,393 target genes (1.7 million interactions), that showed a high predictive power to correctly infer gene functions for TFs involved in a variety of biological processes. The iGRN predicted known functional annotations for 681 TFs and new gene functions for 268 unknown TFs. For regulators predicted to be involved in reactive oxygen species (ROS) stress regulation, 75% were confirmed having a function in ROS and/or physiological stress responses. This included 13 novel ROS regulators that were experimentally validated through ROS-specific phenotypic assays of LOF and gain of function (GOF) lines. The authors observed that the iGRN, leveraging a diverse set of complementary experimental datasets, enabled the identification of new regulators that would have been missed when solely relying on (differential) expression information, demonstrating the power of machine learning approaches.

Next to GRN inference, machine learning can also be applied to directly assign genes to functional categories. For example, Moore and colleagues trained a random forest (RF; reviewed in [43]) model to integrate data spanning five main categories (known function, expression, network, evolution and duplication) and identified 1,220 novel enzymatic genes related to plant specialized metabolism, important for niche-specific interactions between a plant and its environment [45]. Likewise, they identified novel regulatory elements in arabidopsis by incorporating gene expression, TF binding sites, *in vitro* protein-DNA interactions and chromatin accessibility, among others, into a machine learning model, to identify novel cis-regulatory elements (CREs; reviewed in [46]) related to the temporal and JA-(in)dependent response to wounding [47]. Moreover, using metabolomic pathway mapping of DE genes, specialized metabolism wound response CREs were identified. To validate their findings, the wound responsive regulatory effect of the top-predicted CRE on its target gene *GER5* was confirmed using targeted mutagenesis. Additionally, TF binding, irrespective of wounding,

for three late responsive top-predicted CREs were validated using complementary *in vitro* and/or *in vivo* experiments. The RF model returned importance scores for each input feature, which allowed to identify the most predictive features contributing to the overall performance. As such, the authors reported that protein-DNA interaction and chromatin accessibility features were not predictive for the transcriptomic wounding response, likely because these data were not captured from wounded tissues. Thus, when modeling a specific biological process in a targeted approach, it is important that for each data type the experimental conditions are compatible and well-aligned with the process under study. On that basis, another recent study reported the use of matched phenotypic and transcriptomic data, sampled from the same individuals, for maize and arabidopsis plants to build an evolutionary informed machine learning model, predicting nitrogen use efficiency gene-to-phenotype associations [48]. Using a handpicked pool of genotypes that exhibits a broad spectrum of variation in nitrogen use efficiency, evolutionary conserved N-response DE genes were selected as features to build a gradient-boosting predictive model, as well as a GRN using GENIE3 [35]. Results from both models were combined to prioritize regulators involved in nitrogen use efficiency, which was confirmed by genetic analysis using LOF lines for eight arabidopsis TFs and one maize TF.

As highlighted above, gene co-expression analysis can aid the prioritization of causal genes situated within trait-associated loci [21]. As an alternative, a RF model called QTL Causal Gene Finder (QTG-Finder) was developed to prioritize causal genes in arabidopsis and rice QTL, combining polymorphism, functional annotation, network and evolutionary features [49]. More recently, an updated version was published that is capable of building models for any species, based on transfer of causality between orthologous genes [50]. To illustrate its potential, a *Setaria viridis* model was built and employed to prioritize candidate genes in a previously described plant height QTL. While sole application of QTG-Finder2 was deemed not discriminative enough, an intersection with genes transcriptionally upregulated in the internode meristem or cell elongation zone revealed a testable shortlist of 13 candidate causal genes.

Overall, these studies show the potential of machine learning models for integration of multi-omics data, as well as the prioritization of candidate regulators or pathway genes. Moreover, it is possible to test the model for feature importance, which can be used to link specific gene features to experimental conditions providing mechanistic insights in the underlying regulatory mechanisms. However, researchers tend to pursue feature-based machine learning methods, whilst the application of artificial neural networks for functional annotation or identification of regulatory interactions in plants remains limited [51], likely due to the lack of sufficient training data [43].

Single-cell omics increases the resolution for molecular gene characterization

In recent years, single-cell transcriptome profiling has quickly gained traction in plant research, with currently over 30 studies publicly available, spanning different plant species. Unsurprisingly, single-cell RNA sequencing (scRNA-Seq) has already allowed for great advances in our understanding of plant biology (reviewed in [52–55]), partly by unraveling novel gene functions. For example, by identifying cell-type specific expression patterns of arabidopsis TMO5/LHW induced genes using scRNA-Seq, Wendrich and colleagues showed that phosphate starvation triggers TMO5/LHW-dependent cytokinin biosynthesis in the vascular cells which in turn moves to the outer epidermis to result in an increase of root hair density to increase phosphate uptake [56].

Moreover, a study in maize profiled the single-cell transcriptome of developing ears, and integrated obtained results with additional omics data [57]. For example, regulatory modules were identified using co-expression of ChIP-Seq targets for ZmHDZIV6 and ZmMADS16, two TFs specifically expressed in the epidermis and floral organs, respectively. Additionally, by examining genomic co-localization of cell-type specific marker genes with SNPs within yield related ear morphology associated loci, the researchers showed that yield traits are preferentially controlled by cell-type marker genes, and identified three likely yield-trait causal genes.

In rice, a scRNA-Seq study identified two genes, *Os01g0934800* and *Os01g0949900*, regulated by the TF OsNAC78, based on correlating single-cell expression profiles and DE analysis, followed by experimental validation using the yeast one-hybrid system [58]. While *OsNAC78* remains to be experimentally characterized, its newly identified target genes were previously described to play a role in detoxification and reactive oxygen species scavenging, respectively, thereby hinting at putative functions of *OsNAC78*.

Another study in *Arachis hypogaea* (peanut) identified 11 TFs involved in the leaf primordia to epidermal tissue transition, using pseudo-time analysis [59]. Further investigation of one candidate, *AHL23*, confirmed increased epidermal cell number upon ectopic expression in arabidopsis, likely through altered phytohormone regulation.

While the above mentioned single-cell technology undeniably increased the resolution in which we can study the transcriptome, any physical context is lost upon dissociation of the tissue into isolated cells, which can only be reconstructed for a select number of well-studied organs such as the arabidopsis root. Conversely, **spatial transcriptomics** retains physical interactions of individual cells, allowing to study each individual cell's gene expression in light of its position in the tissue [60]. Currently, cellular resolution cannot always be achieved when using untargeted transcriptome-wide methods [61,62], while targeted methods can profile a predefined set of transcripts at a given subcellular localization [63,64].

In plants, untargeted spatial transcriptomics has been applied in arabidopsis, *Populus tremula*, and *Picea abies* tissues, recovering known and novel biology [62]. For example, *P. tremula* orthologs of arabidopsis developmental regulators were upregulated in developing buds compared to dormant buds. Additionally, the authors developed a linear model to quantify DE that includes components representing biological and technical replicates, sequencing array spots and defined tissue domains within each assayed section. This model identified significant differences in tissue domain specific genes and pathways, such as the upregulation of the stamen filament development pathway at the

site of stamen filament elongation, as well as the pollen exine formation pathway, needed for formation of the pollen wall, in the tapetum layer.

Moreover, spatial information can be leveraged to expose the function of expressed genes, as illustrated by Laureyns and co-workers [64]. In this study, the authors applied spatial transcriptomics to colocalize the transcripts of 90 genes in the maize shoot apical meristem, and thereby inferring a role of *PLA1* in marking boundaries between indeterminate cells and determinate lateral organ primordia.

To conclude, these studies illustrate the potential of single cell profiling for functional characterization. Moreover, spatial transcriptomics can uncover gene function by analyzing transcript abundance in a spatial, subcellular context. However, current efforts have focused on specific genes and co-expression relationships, while transcriptome-wide single-cell co-expression analysis and/or GRN inference have not been thoroughly explored in plants.

Concluding remarks and future perspectives

The emergence of multi-omics and single-cell profiling is rapidly changing the way how molecular profiling and systems biology approaches are being used to unravel signaling cascades controlling plant biology. While technological improvements continue to increase the number of genes that can be studied, as well as the cellular resolution, new methods supporting integrative data analysis are being developed in parallel.

An important area in plant research that remains to be explored is co-expression network analysis and GRN inference using single-cell data sets. Such approaches have the potential to transcend their bulk tissue counterparts and provide new functional information at an unprecedented resolution. For example, gene co-expression signals present in bulk tissue transcriptome data might be diluted by the presence of several functionally and morphologically distinct cell-types, each with its own transcriptomic fingerprint [65], within a single sample. Therefore, obtaining gene expression profiles

from individual cells and analyzing co-expression relationships among genes over all profiled cells might reveal novel functional insights [53,66]. Similarly, analyzing cell-type clusters separately might expose rare co-expression, and possibly functional, relationships unique to a specific cell-type, as suggested for the mouse brain [66]. However, co-expression analysis of single-cell data is still limited by poor co-expression signals within individual data sets, possibly due to the abundance of drop-outs and the overall low fraction of the transcriptome that is captured [67].

The coordinated activity of TFs, controlling cell-type specific gene expression, is a key driver of cell identity [65] and it has been hypothesized that cell-type specific GRNs can be mined to identify key molecular players and explore their distinct cell-type specific functions [68]. For example, a TF might regulate a set of functionally related target genes in a certain cell-type, exposing its function in that cell-type, while in another cell-type that TF can regulate another set of target genes involved in a different process. Thus, methods designed to delineate cell-type specific GRNs using scRNA-Seq data [69–71] present a new way of analyzing fine-grained gene function and regulation.

While different omics profiling methods offer a detailed molecular read-out about genes' activity and regulation, most of the recent multi-omics studies performed downstream experimental validation of a handful of genes. A major challenge is how the information obtained through omics experiments can be used to functionally annotate a larger number of genes in a high-throughput manner. While the experimental PlantGSAD gene sets (Box 1), collected through manual curation of e.g. supplemental tables reporting responsive or functional genes, offer one way to do this, automated procedures to extract and leverage functional gene sets from single-cell genomics experiments are currently lacking. Clearly, the development of efficient data analysis methods will further boost the characterization of the functional gene landscape in plant biology.

Outstanding questions

- When starting from a targeted multi-omics profiling experiment, how can we increase the number of genes that are modeled, predicted and validated?
- For untargeted multi-omics based gene function inference, how to define the balance between the number of genes that are assigned new gene functions and the specificity and accuracy of these novel functional annotations?
- What is the ideal combination of different omics data types to unravel gene functions for specific biological processes or pathways of interest?
- Can current gene function inference strategies be applied to or adapted for single-cell profiling data, or are new algorithms needed to fully exploit the unique properties of these new data types?
- How can the ever increasing resolution of untargeted spatial transcriptomics, as well as the prospect of other spatially resolved omics data, be fully leveraged to learn gene function?
- How can multi-omics profiling and integrative data analysis be used to characterize gene functions for other gene types, such as long noncoding RNAs?
- When studying a high-level stimulus with complex regulatory wiring and affecting diverse phenotypic traits, how to prioritize the phenotypes to score in a reverse genetics experiment?

Box 1. Functional classification systems and annotation resources

The function of a single gene can be described in many different ways, and various descriptions can point towards the same function. This can result in interpretation errors and is especially problematic in the context of querying databases, automation and data mining. To alleviate this problem, functions can be formally described using ontologies or controlled vocabularies, such as the **Gene Ontology (GO)** [94]. While GO is the most widely used, other vocabularies such as the **Plant Ontology (PO)** and the **Plant Trait Ontology (TO)** capture other types of functional information and follow the same design principles as GO [95]. **Evidence codes**, explicit to any ontology-based annotation, denote the support for each annotation, allowing researchers to attribute a level of confidence to a specific annotation. Likewise, the proposed GO Annotation File (GAF2) format requires several other important data fields, such as a reference to the source of the annotation [96], which improves traceability. Next to the above mentioned ontologies, other systems exist which focus on metabolic pathways and have been applied to or designed specifically for plant species. For example, the Plant Metabolic Network (PMN) is a resource of species-specific metabolism databases for 126 plants and algae, combining experimentally-supported and computationally-predicted information on metabolites, reactions, enzymes and pathways [97]. Moreover, the PMN includes a multi-species database PlantCyc that contains experimental metabolic data for 519 plant species, as well as a selection of tools for integration with additional omics data. Likewise, MapMan enables users to visualize omics data onto plant pathways, which are defined in the hierarchical MapMan BIN ontology and encompass metabolic and regulatory processes, signaling pathways and stress responses [98]. A frequently used alternative not specific to plant species is the Kyoto Encyclopedia of Genes and Genomes (KEGG), which serves as a system to link genes to molecular functions through grouping of orthologous genes, and currently contains 129 plant species [99]. A large collection of online plant functional genomics resources and databases exists, often collecting data for multiple species and different omics technologies. One prominent example is the Plant Gene Set Annotation Database (PlantGSAD), where researchers have collected contents from multiple public databases, compiling a resource covering nine main annotation categories (such as GO/PO/TO annotations and KEGG/PlantCyc/MapMan pathway annotations) and 44 plant species [100]. However, for ontology annotations, evidence codes are not reported and can only be queried from the source database, hindering automated analyses.

Box 2. Integration strategies for multi-omics data

Exploiting multiple types of omics data can be approached in different ways. Integration strategies applied by the studies discussed in the current review can be classified into four main categories. First, **conceptual integration** [22] is a basic, yet effective strategy that relies on separate handling of individual data sets, followed by matching or overlaying the results. For example, overlap analysis of gene sets derived from different omics layers [31–33,36,50], a comparison of an omics feature (e.g. chromatin accessibility or genomic colocalization) between gene modules derived from another data type [21,37,38], as well as joint analysis of individually delineated networks [38,40] are commonly applied. Second, **statistical integration** [22,23] aims to identify statistical associations between genes and/or other elements profiled using different omics methods. For example, correlation-based integration quantifies relationships between individual molecules (e.g. transcripts and proteins) from different omics layers [31–33], whereas multivariate-based integration can identify trends, such as covariance associations, across multiple omics layers [75]. Data set concatenation followed by unsupervised clustering allows direct grouping of genes based on different data types [33]. A related frequently applied strategy is pathway-based integration, where different omics data sets are mapped onto existing pathway databases, often using enrichment statistics [31–33,36,40,45,47]. Third, **network-based integration** [24] aims to connect molecular entities into a single biological network based on associations originating different omics layers. A popular approach is to take the union or intersection of individual omics-specific networks to produce a merged network [31–33,40]. Alternatively, clustering genes in modules based on different omics layers prior to network inference can directly produce an integrated network [33]. Finally, **machine learning** models allow integration of various omics features [101], such as genomic properties and expression levels, directly coupled with regulatory or functional classification of genes. For example, multiple omics layers can be used as input features for a model predicting traits or regulatory interactions [44,45,49,50], or one omics data type can be used as input feature while another omics layer is used to define the output class [47,48].

References

1 Peters, J.L. *et al.* (2003) Forward genetics and map-based cloning approaches. *Trends in Plant Science* 8, 484–491

2 Rhee, S.Y. and Mutwil, M. (2014) Towards revealing the functions of all genes in plants. *Trends in Plant Science* 19, 212–221

3 Gilchrist, E. and Haughn, G. (2010) Reverse genetics techniques: engineering loss and gain of gene function in plants. *Briefings in Functional Genomics* 9, 103–110

4 Somerville, C. and Dangl, J. (2000) Plant Biology in 2010. *Science* 290, 2077–2078

5 Provart, N.J. *et al.* (2016) 50 years of Arabidopsis research: highlights and future directions. *New Phytologist* 209, 921–944

6 Hirschi, K.D. (2003) Insertional mutants: a foundation for assessing gene function. *Trends in Plant Science* 8, 205–207

7 Briggs, G.C. *et al.* (2006) Unequal genetic redundancies in Arabidopsis – a neglected phenomenon? *Trends in Plant Science* 11, 492–498

8 Kim, E.-D. and Sung, S. (2012) Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends in Plant Science* 17, 16–21

9 Ariel, F. *et al.* (2015) Battles and hijacks: noncoding transcription in plants. *Trends in Plant Science* 20, 362–371

10 He, M. *et al.* (2022) MicroRNAs: emerging regulators in horticultural crops. *Trends in Plant Science* 0,

11 Mishra, B. *et al.* (2019) Systems Biology and Machine Learning in Plant–Pathogen Interactions. *MPMI* 32, 45–55

12 Clark, N.M. *et al.* (2022) To the proteome and beyond: advances in single-cell omics profiling for plant systems. *Plant Physiology* 188, 726–737

13 Bartlett, A. *et al.* (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat Protoc* 12, 1659–1672

14 Gaudinier, A. and Brady, S.M. (2016) Mapping Transcriptional Networks in Plants: Data-Driven Discovery of Novel Biological Mechanisms. *Annu Rev Plant Biol* 67, 575–594

15 Bolduc, N. *et al.* (2012) Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes Dev.* 26, 1685–1690

16 Chang, K.N. *et al.* (2013) Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. *eLife* 2, e00675

17 Heyndrickx, K.S. *et al.* (2014) A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. *Plant Cell* 26, 3894–3910

18 Kulkarni, S.R. *et al.* (2018) TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Res* 46, e31

19 Gore, M.A. *et al.* (2009) A first-generation haplotype map of maize. *Science* 326, 1115–1117

20 Ramstein, G.P. *et al.* (2019) Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theor Appl Genet* 132, 559–567

21 Schaefer, R.J. *et al.* (2018) Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. *The Plant Cell* 30, 2922–2942

22 Cavill, R. *et al.* (2016) Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics* 17, 891–901

23 Jamil, I.N. *et al.* (2020) Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology. *Frontiers in Plant Science* 11,

24 Zhou, G. *et al.* (2020) Network-Based Approaches for Multi-omics Integration. In *Computational Methods and Data Analysis for Metabolomics* (Li, S., ed), pp. 469–487, Springer US

25 Treves, H. *et al.* (2020) Multi-omics reveals mechanisms of total resistance to extreme illumination of a desert alga. *Nat. Plants* 6, 1031–1043
26 Reynoso, M.A. *et al.* (2022) Gene regulatory networks shape developmental plasticity of root cell types under water extremes in rice. *Developmental Cell* 0,
27 Zhang, X. *et al.* (2022) Integrated multi-omic data and analyses reveal the pathways underlying key ornamental traits in carnation flowers. *Plant Biotechnology Journal* n/a,
28 Wang, Z. *et al.* (2022) Transcriptome Co-expression Network and Metabolome Analysis Identifies Key Genes and Regulators of Proanthocyanidins Biosynthesis in Brown Cotton. *Frontiers in Plant Science* 12,
29 Ding, X. *et al.* (2022) Microautophagy Mediates Vacuolar Delivery of Storage Proteins in Maize Aleurone Cells. *Frontiers in Plant Science* 13,
30 Wang, W.-Q. *et al.* (2022) A multiomic study uncovers a bZIP23-PER1A-mediated detoxification pathway to enhance seed vigor in rice. *Proc Natl Acad Sci U S A* 119, e2026355119
31 Zander, M. *et al.* (2020) Integrated multi-omics framework of the plant response to jasmonic acid. *Nature Plants* 6, 290–302
32 Walley, J.W. *et al.* (2016) Integration of omic networks in a developmental atlas of maize. *Science* 353, 814–818
33 Clark, N.M. *et al.* (2021) Integrated omics networks reveal the temporal signaling events of brassinosteroid response in Arabidopsis. *Nat Commun* 12, 5858
34 Clark, N.M. *et al.* (2019) Stem-cell-ubiquitous genes spatiotemporally coordinate division through regulation of stem-cell-specific gene networks. *Nat Commun* 10, 5574
35 Huynh-Thu, V.A. *et al.* (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE* 5, e12776
36 Wang, G. *et al.* (2021) Cross species multi-omics reveals cell wall sequestration and elevated global transcript abundance as mechanisms of boron tolerance in plants. *New Phytologist* 230, 1985–2000
37 Reynoso, M.A. *et al.* (2019) Evolutionary flexibility in flooding response circuitry in angiosperms. *Science* 365, 1291–1295
38 Zhou, P. *et al.* (2020) Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions. *The Plant Cell* 32, 1377–1396
39 Serin, E.A.R. *et al.* (2016) Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in Plant Science* 7,
40 Depuydt, T. and Vandepoele, K. (2021) Multi-omics network-based functional annotation of unknown Arabidopsis genes. *The Plant Journal* 108, 1193–1212
41 Rao, X. and Dixon, R.A. (2019) Co-expression networks for plant biology: why and how. *Acta Biochimica et Biophysica Sinica* 51, 981–988
42 Lee, T. *et al.* (2015) AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Res* 43, D996-1002
43 Mahood, E.H. *et al.* (2020) Machine learning: A powerful tool for gene function prediction in plants. *Appl Plant Sci* 8, e11376
44 De Clercq, I. *et al.* (2021) Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators. *Nature Plants* 7, 500–513
45 Moore, B.M. *et al.* (2019) Robust predictions of specialized metabolism genes through machine learning. *PNAS* 116, 2344–2353
46 Schmitz, R.J. *et al.* (2021) Cis-regulatory sequences in plants: their importance, discovery, and future challenges. *Plant Cell* DOI: 10.1093/plcell/koab281
47 Moore, B.M. *et al.* (2021) Modeling temporal and hormonal regulation of plant transcriptional response to wounding. *The Plant Cell* DOI: 10.1093/plcell/koab287

48 Cheng, C.-Y. *et al.* (2021) Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. *Nat Commun* 12, 5627
49 Lin, F. *et al.* (2019) QTG-Finder: A Machine-Learning Based Algorithm To Prioritize Causal Genes of Quantitative Trait Loci in Arabidopsis and Rice. *G3 (Bethesda)* 9, 3129–3138
50 Lin, F. *et al.* (2020) QTG-Finder2: A Generalized Machine-Learning Algorithm for Prioritizing QTL Causal Genes in Plants. *G3 Genes|Genomes|Genetics* 10, 2411–2421
51 Washburn, J.D. *et al.* (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences* 116, 5542–5549
52 Rich-Griffin, C. *et al.* (2020) Single-Cell Transcriptomics: A High-Resolution Avenue for Plant Functional Genomics. *Trends in Plant Science* 25, 186–197
53 Seyfferth, C. *et al.* (2021) Advances and Opportunities in Single-Cell Transcriptomics for Plant Research. *Annu. Rev. Plant Biol.* 72, 847–866
54 Shaw, R. *et al.* (2021) Single-Cell Transcriptome Analysis in Plants: Advances and Challenges. *Mol Plant* 14, 115–126
55 Minne, M. *et al.* (2022) Advancing root developmental research through single-cell technologies. *Current Opinion in Plant Biology* 65, 102113
56 Wendrich, J.R. *et al.* (2020) Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions. *Science* 370, eaay4970
57 Xu, X. *et al.* (2021) Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. *Developmental Cell* 56, 557-568.e6
58 Xie, Y. *et al.* (2020) Single-Cell RNA Sequencing Efficiently Predicts Transcription Factor Targets in Plants. *Frontiers in Plant Science* 11,
59 Liu, H. *et al.* (2021) Single-cell RNA-seq describes the transcriptome landscape and identifies critical transcription factors in the leaf blade of the allotetraploid peanut (Arachis hypogaea L.). *Plant Biotechnology Journal* 19, 2261–2276
60 Giacomello, S. (2021) A new era for plant science: spatial single-cell transcriptomics. *Current Opinion in Plant Biology* 60, 102041
61 Ståhl, P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82
62 Giacomello, S. *et al.* (2017) Spatially resolved transcriptome profiling in model plant species. *Nature Plants* 3, 1–11
63 Ke, R. *et al.* (2013) In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* 10, 857–860
64 Laureyns, R. *et al.* (2022) An in situ sequencing approach maps PLASTOCHRON1 at the boundary between indeterminate and determinate cells. *Plant Physiology* 188, 782–794
65 Arendt, D. *et al.* (2016) The origin and evolution of cell types. *Nat Rev Genet* 17, 744–757
66 Harris, B.D. *et al.* (2021) Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. *Cell Systems* 12, 748-756.e3
67 Crow, M. and Gillis, J. (2018) Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? *Trends Genet* 34, 823–831
68 Tripathi, R.K. and Wilkins, O. (2021) Single cell gene regulatory networks in plants: Opportunities for enhancing climate change stress resilience. *Plant, Cell & Environment* 44, 2006–2017
69 Pratapa, A. *et al.* (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 17, 147–154
70 Van de Sande, B. *et al.* (2020) A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* 15, 2247–2276

71 Skok Gibbs, C. *et al.* (2022) High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. *Bioinformatics* 38, 2519–2528
72 Proost, S. *et al.* (2009) PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. *The Plant Cell* 21, 3718–3731
73 Bolger, M.E. *et al.* (2018) Plant genome and transcriptome annotations: from misconceptions to simple solutions. *Brief Bioinform* 19, 437–449
74 Wimalanathan, K. *et al.* (2018) Maize GO Annotation—Methods, Evaluation, and Review (maize-GAMER). *Plant Direct* 2, e00052
75 de Abreu e Lima, F. *et al.* (2018) Unraveling lipid metabolism in maize with time-resolved multi-omics data. *The Plant Journal* 93, 1102–1115
76 Nguyen, K.L. *et al.* (2019) Next-Generation Sequencing Accelerates Crop Gene Discovery. *Trends in Plant Science* 24, 263–274
77 Varshney, R.K. *et al.* (2021) Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends in Plant Science* 26, 631–649
78 Kilian, J. *et al.* (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal* 50, 347–363
79 Klepikova, A.V. *et al.* (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *The Plant Journal* 88, 1058–1070
80 Schaefer, R.J. *et al.* (2017) Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1860, 53–63
81 Haque, S. *et al.* (2019) Computational prediction of gene regulatory networks in plant growth and development. *Current Opinion in Plant Biology* 47, 96–105
82 Dahhan, D.A. *et al.* (2022) Proteomic characterization of isolated Arabidopsis clathrin-coated vesicles reveals evolutionarily conserved and plant-specific components. *The Plant Cell* DOI: 10.1093/plcell/koac071
83 Mergner, J. and Kuster, B. (2022) Plant Proteome Dynamics. *Annu Rev Plant Biol* DOI: 10.1146/annurev-arplant-102620-031308
84 Millar, A.H. *et al.* (2019) The Scope, Functions, and Dynamics of Posttranslational Protein Modifications. *Annu Rev Plant Biol* 70, 119–151
85 Hall, R.D. *et al.* (2022) High-throughput plant phenotyping: a role for metabolomics? *Trends in Plant Science* 0,
86 Gallusci, P. *et al.* (2017) Epigenetics for Plant Improvement: Current Knowledge and Modeling Avenues. *Trends in Plant Science* 22, 610–623
87 Crisp, P.A. *et al.* (2020) Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *PNAS* 117, 23991–24000
88 Kulkarni, S.R. *et al.* (2019) Enhanced Maps of Transcription Factor Binding Sites Improve Regulatory Networks Learned from Accessible Chromatin Data1[OPEN]. *Plant Physiol* 181, 412–425
89 Taylor-Teeples, M. *et al.* (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature* 517, 571–575
90 Tu, X. *et al.* (2020) Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications* 11, 5089
91 Struk, S. *et al.* (2019) Exploring the protein–protein interaction landscape in plants. *Plant, Cell & Environment* 42, 387–409
92 Ben-Amar, A. *et al.* (2016) Reverse Genetics and High Throughput Sequencing Methodologies for Plant Functional Genomics. *Curr Genomics* 17, 460–475

93 Sharma, N. *et al.* (2022) Advances in potato functional genomics: implications for crop improvement. *Plant Cell Tiss Organ Cult* 148, 447–464
94 Ashburner, M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25–29
95 Cooper, L. *et al.* (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research* 46, D1168–D1180
96 Balakrishnan, R. *et al.* (2013) A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford)* 2013, bat054
97 Hawkins, C. *et al.* (2021) Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *Journal of Integrative Plant Biology* 63, 1888–1905
98 Schwacke, R. *et al.* (2019) MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Molecular Plant* 12, 879–892
99 Kanehisa, M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44, D457–D462
100 Ma, X. *et al.* (2022) PlantGSAD: a comprehensive gene set annotation database for plant species. *Nucleic Acids Research* 50, D1456–D1467
101 Reel, P.S. *et al.* (2021) Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances* 49, 107739

Figure

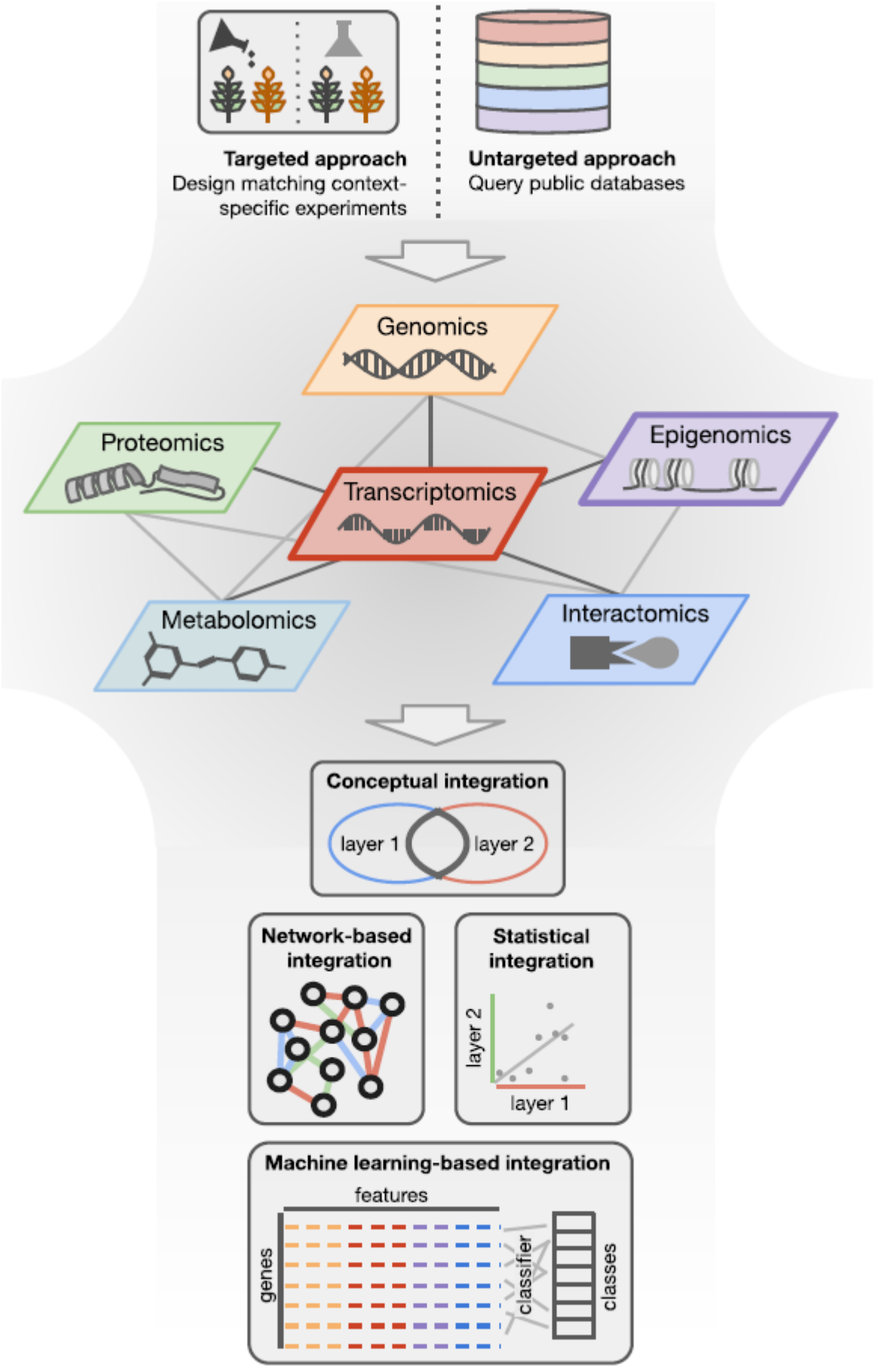
Targeted approach
Design matching context-specific experiments
Untargeted approach
Query public databases
Genomics
Proteomics
Transcriptomics
Epigenomics
Metabolomics
Interactomics
Conceptual integration
layer 1
layer 2
Network-based integration
Statistical integration
layer 2
layer 1
Machine learning-based integration
features
genes
classifier
classes

Figure 1. Multi-omics strategies for charting gene functions in plants. Multi-omics data can be obtained from experimental procedures or queried from functional genomics databases. Transcriptomics is often a central piece of the study design (connections highlighted with heavy lines), and each additional omics layer can expose a complementary aspect of gene function. Moreover, different combinations can be used to answer specific biological questions. Omics data types where single cell resolution has emerged in plant science are highlighted with a heavy border. Multiple integration strategies exist to exploit the information captured by the various omics layers and allow the generation of new testable hypotheses for biological processes and pathways.

Tables

Table 1. Selection of established methods for functional genomics.

Type	Technique	Application in functional annotation	Strength	Weakness	Selected references
Genomics	Homology-based	Transfer of functions through sequence-similarity to experimentally characterized genes	Fast and widely applicable for any species with sequenced genome	Gene families / protein domains without experimentally characterized members are excluded; functional divergence of paralogs	[72–74]
	Quantitative trait loci linkage mapping	Identification of possible causal genes in associated loci	Powerful statistical frameworks that uncover the genetic architecture of complex traits	Non-trivial identification of causal genes	[75,76]
	Genome-wide association studies				[21,77]
Transcriptomics	Differential expression	Identification of genes with altered condition-specific expression profile	Apparent gene-to-condition association	Limited correlation between protein and mRNA levels	[78,79]
	Co-expression networks	Guilt-by-association: functional identification of network modules	Transcriptome-wide biological interpretation of gene expression	Many non-functional edges	[40,80]
	Expression-based network inference methods[a]	Guilt-by-association: functional identification of regulatory transcription factor – target gene interactions	Interrogation of transcriptome-wide gene function in a regulatory context		[44,81]
Proteomics	Differential expression	Identification of genes with altered condition-specific protein abundance profile	Profiling of functional entities	Expensive experimental profiling; lower sensitivity compared to other omics profiling methods	[82,83]
	Post-translational modifications	Identification of gene product molecular functional characteristics	Higher resolution into molecular function		[84]
Metabolomics	Differential abundance	Association with differential gene expression profiles	Highly relevant aspect of the plant phenotype	Expensive experimental profiling	[85]
Epigenomics	DNA methylation	Identification of functional and regulatory sequences	Particularly relevant for plant species with large genomes and a high fraction of intergenic space	Low variability across developmental stages and environmental stresses limits functional annotation	[86,87]
	Open chromatin profiling	Differential accessibility of regulatory sequences	Variable between conditions and tissues	Learning specific regulatory interactions is challenging	[46,88]
Interactomics	Experimental protein-DNA interaction networks	Guilt-by-association: functional identification of regulatory target modules	Interrogation of transcription factor function in a regulatory context	Many non-functional interactions	[14,89,90]
	Experimental protein-protein interaction networks	Guilt-by-association: functional identification of network modules	Strong implication of functional relation between physical interactors	Need for quality control and replicates to identify reliable interactions	[91]
Reverse genetics	Gain-of-function (GOF)	Investigation of altered phenotype between wild-type and mutant plants	Heterologous expression systems suitable for non-model species	Time consuming, complex phenotypes	[92,93]
	Loss-of-function (LOF)		Many efficient techniques available	Phenotypes obscured by functional redundancy	

[a] While less established, also other omics data types can be used to infer networks.