the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2021) **108,** 1585–1596 doi: 10.1111/tpj.15547

FOCUSED REVIEW

Arabidopsis bioinformatics: tools and strategies

Alex Cantó-Pastor[1,†], G. Alex Mason[1,†], Siobhan M. Brady[1] and Nicholas J. Provart[2]*

[1]*Department of Plant Biology and Genome Center, University of California Davis, 1 Shields Avenue, Davis, CA 95616, USA, and*

[2]*Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada*

Received 21 June 2021; revised 1 October 2021; accepted 19 October 2021; published online 9 November 2021.
*For correspondence (e-mail: nicholas.provart@utoronto.ca).
[†]These authors contributed equally to this work.

SUMMARY

The sequencing of the *Arabidopsis thaliana* genome 21 years ago ushered in the genomics era for plant research. Since then, an incredible variety of bioinformatic tools permit easy access to large repositories of genomic, transcriptomic, proteomic, epigenomic and other '-omic' data. In this review, we cover some more recent tools (and highlight the 'classics') for exploring such data in order to help formulate quality, testable hypotheses, often without having to generate new experimental data. We cover tools for examining gene expression and co-expression patterns, undertaking promoter analyses and gene set enrichment analyses, and exploring protein–protein and protein–DNA interactions. We will touch on tools that integrate different data sets at the end of the article.

Keywords: transcriptomics, hypothesis generation, bioinformatics, protein–protein interactions, co-expression, functional genomics.

Summary

- The use of high-throughput "-omic" technologies has dramatically increased over the last several years, and most data sets generated have been made publicly available after publication.
- Several web tools have been developed that allow researchers to query and explore these public data sets, and validate or generate hypotheses for their genes of interest using them.
- From phylogenomics to protein-protein interaction networks, we review some of the most relevant and up-to-date bioinformatic tools available online.
- Using these tools, integrative analyses across multiple types of data can provide insights and will help with elucidating the role of a gene or set of genes of interest.

INTRODUCTION

Our understanding of plant biology has accelerated in the past decade as cheaper and more processive methods for sequencing nucleic acids have been developed. These have enabled high-throughput analyses of genomes, epigenomes, transcriptomes and protein–protein or protein–DNA interactions (for an excellent overview of these technologies, see Reuter et al., 2015). Quantitative proteomic and metabolomic detection methods have enjoyed a similar boost. Data generated using such methods obviously help answer a plant biologist's research question. What is less obvious is that once these data sets are made publicly available they are useful to other plant biologists to explore and answer their own biological questions (Chory et al., 2000). These data sets can be queried to help design experiments or generate hypotheses at the click of a mouse. Such hypotheses can be followed up in the lab with molecular methods. This review provides an overview

of useful (mostly) web-based tools for experimental researchers.

Here, we will touch on well-cited web-based tools that bring together data from several sources – these are often more useful to the typical Arabidopsis researcher than single data investigator-run databases. We will focus on tools introduced in the past 5 years and on those that are updated regularly. The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org) is updated constantly and we also direct you to a Current Protocols in Bioinformatics overview of using this sequence-centric Arabidopsis database for finding information about genes (Reiser et al., 2017). ARAPORT (https://www.araport.org; Krishnakumar et al., 2015) and SIGnAL (http://signal.salk.edu; Alonso et al., 2003) are two additional websites for exploring sequences and identifying transposon insertions in your gene(s) of interest; we will examine these in brief, along with websites associated with the 1001 Genomes Project (https://1001genomes.org).

We will start with a discussion about the tools used for querying transcriptome data sets, the most comprehensive of all the 'omic' data types, and how to query these data sets in both targeted and correlative ways using co-expression analyses. We will touch on several newer tools for querying single-cell RNA-seq data. Querying gene expression data can be useful for focusing the search for mutant phenotypes or for generating candidates that have novel genetic associations with a given biological process. We will also highlight protein–protein interaction tools for Arabidopsis and tools for performing promoter analyses.

The tools covered in this review are listed in Table 1 and are illustrated in Figure 1. Step-by-step instructions for many of these tools are described in a recent chapter by Mason et al. (2021). Brady and Provart (2009) wrote a review article on bioinformatic tools for hypothesis generation that remains relevant and is still well worth reading. Furthermore, TAIR maintains a 'super-portal' to keep track of and categorize various Arabidopsis tools (https://conf.arabidopsis.org/display/COM/Resources). Lastly, one of the co-authors of this review has created a course called Plant Bioinformatics (https://www.coursera.org/learn/plant-bioinformatics/) that you can evaluate for free. Many of the tools described herein are explored in this online lab course.

GENOME DATABASES

As mentioned, TAIR curators have put together a great how-to guide for their website, one of the most widely used Arabidopsis portals (Reiser et al., 2017). TAIR curators are constantly updating functional annotations based on new publications. Another starting place for information about Arabidopsis genes is ARAPORT's ThaleMine, now maintained by the Bio-Analytic Resource for Plant Biology (BAR; Pasha et al., 2020). ThaleMine collects data from published literature and data sets and provides computational and visualization tools (and programmatic access if you're a computational biologist, through BlueGenes; http://bluegenes.apps.intermine.org/thalemine). Lastly, the 1001 Genomes Project provides data on more than 1001 ecotypes (also called strains, accessions or genotypes) of *Arabidopsis thaliana* with a wide geographic distribution (1001 Genomes Consortium, 2016). With the PolyMorph tool at https://1001genomes.org, synonymous or non-synonymous variants in the sequence of a gene may be identified, as can small insertions or deletions. We also recommend checking out two related and highly complementary tools associated with the 1001 Genomes Project: AraGWAS and AraPheno (Togninalli et al., 2020). Both have elegantly designed interfaces for efficiently exploring the data from the 1001 Genomes Project in the context of genome-wide association studies and phenotypes. The AraPheno tool permits the exploration of more than 462 phenotypes across 1496 accessions.

Pre-computed gene trees and other genomic resources

As a result of the number of available sequenced plant genomes, phylogenomics is becoming more valuable for understanding gene function in related species. Evolutionary relationships of a gene can be used to 'lift over' annotations from homologs to help predict function (Andrade et al., 1999). Gene duplication events that might affect reverse genetic strategies (e.g. generating a double-knockout mutant) are readily apparent through phylogenetic analyses. Hypotheses that can be generated for a gene of interest could include whether subfunctionalization of the duplicates has occurred or whether there is redundancy in function. For translational researchers working in other plant species, identifying the Arabidopsis homolog most likely to be the ortholog in that species might be useful. Pre-computed phylogenetic trees, from Ensembl Plants, PLAZA or PANTHER, are useful for all of these purposes. Ensembl Plants provides genome-scale data from plants (Kersey et al., 2018) and is maintained by the European Bioinformatics Institute and curators at Gramene. Within these trees, one can collapse and expand branches in the tree by clicking on the squares (nodes) and triangles (subtrees) to achieve the desired display, and to make it readily apparent whether a gene duplication event has occurred in a given species or lineage. Nodes also contain bootstrap values. PLAZA aggregates genomic data produced by different genome sequencing initiatives in a comparative genomics framework (Van Bel et al., 2018). With PLAZA 4.0 it is possible to create a custom tree using the Interactive Phylogenetics Module in its toolbox to be able to focus on species/homologs of interest. The PANTHER platform, which is updated monthly, provides a suite of tools and data for studying evolutionary relationships, gene function, biochemical pathways and other data (Mi

Table 1 Tools, URLs and references

Methods	Tool	Web	Ref.
Genome databases	TAIR	http://arabidopsis.org	(Reiser et al., 2017)
	ARAPORT	https://www.araport.org/	(Krishnakumar et al., 2015)
	1001 Genomes Project	https://1001genomes.org/ https://arapheno.1001genomes.org/ https://aragwas.1001genomes.org/	(1001 Genomes Consortium, 2016; Togninalli et al., 2020)
Pre-computed gene trees and other genome resources	Ensembl Plants	https://plants.ensembl.org/index.html	(Kersey et al., 2018)
	PLAZA	https://bioinformatics.psb.ugent.be/plaza/	(Van Bel et al., 2018)
	PANTHER	http://www.pantherdb.org/	(Mi et al., 2010)
Epigenomic tools	EPIC-CoGe	https://genomevolution.org/CoGe/GenomeView.pl	(Nelson et al., 2018)
Expression analysis	eFP Browser/eFP-Seq Browser	http://bar.utoronto.ca; http://bar.utoronto.ca/eFP-Seq_Browser/	(Sullivan et al., 2019; Winter et al., 2007)
	GENEVESTIGATOR	http://www.genevestigator.com/; https://genevisible.com/search	(Hruz et al., 2008)
	TravaDB	http://travadb.org	(Klepikova et al., 2016)
Co-expression tools	ATTED II	http://atted.jp	(Aoki et al., 2016)
	Expression Angler	http://bar.utoronto.ca	(Toufighi et al., 2005)
	AraNet	https://www.inetbio.org/aranet	(Lee et al., 2010)
	AtCAST	http://atpbsmd.yokohama-cu.ac.jp/cgi/atcast/home.cgi	(Kakei and Shimada, 2015)
Promoter analysis	Cistome	http://bar.utoronto.ca/cistome/cgi-bin/BAR_Cistome.cgi	(Austin et al., 2016)
	ePlant	http://bar.utoronto.ca/eplant	(Waese et al., 2017)
	MEME: FIMO and AME	http://meme-suite.org/	(Grant et al., 2011; McLeay and Bailey, 2010)
GO/functional enrichment analyses	AgriGO	http://systemsbiology.cau.edu.cn/agriGOv2/	(Tian et al., 2017)
	MAPMAN	https://mapman.gabipd.org/	(Schwacke et al., 2019; Thimm et al., 2004)
	PLAZA	https://bioinformatics.psb.ugent.be/plaza/	(Van Bel et al., 2018)
	BiNGO	https://apps.cytoscape.org/	(Maere et al., 2005)
	AmiGO	http://amigo.geneontology.org/rte	(Carbon et al., 2009)
Pathway visualization	AraCyc	www.plantcyc.org/	(Mueller et al., 2003)
	MAPMAN	https://mapman.gabipd.org/	(Schwacke et al., 2019; Thimm et al., 2004)
Protein information	SUBA Live	http://suba.live/	(Hooper et al., 2017)
	Cell eFP Browser	http://bar.utoronto.ca/cell_efp/cgi-bin/cell_efp.cgi	(Winter et al., 2007)
	P^3DB	https://p3db.org/home	(Gao et al., 2009; Yao et al., 2012, 2014)
	Plant PTM Viewer	https://dev.bits.vib.be/ptm-viewer/index.php	(Willems et al., 2019)
	Ubiquitination Site tool	http://bioinformatics.psb.ugent.be/webtools/ubiquitin_viewer/	(Walton et al., 2016)
	ATHENA	http://athena.proteomics.wzw.tum.de/	(Mergner et al., 2020)
Protein–protein interaction	Arabidopsis Interactions Viewer 2	http://bar.utoronto.ca/interactions2/	(Dong et al., 2019)
	BioGRID	http://thebiogrid.org	(Chatr-Aryamontri et al., 2017)
Integrated tools	TF2Network	http://bioinformatics.psb.ugent.be/webtools/TF2Network/	(Kulkarni et al., 2018)
	GeneMANIA	http://genemania.org/	(Warde-Farley et al., 2010)
	CORNET	https://bioinformatics.psb.ugent.be/cornet/	(Van Bel and Coppens, 2017)
	ePlant	http://bar.utoronto.ca	(Waese et al., 2017)
Targeting tools	CRISPR-PLANT	https://www.genome.arizona.edu/crispr/CRISPRsearch.html	(Xie et al., 2014)
	CRISPR-P	http://crispr.hzau.edu.cn/CRISPR2/	(Lei et al., 2014; Liu et al., 2017)
	WMD3	http://wmd3.weigelworld.org	(Ossowski et al., 2008)
	SIGnAL T-DNA Express	http://signal.salk.edu/	(Alonso et al., 2003)

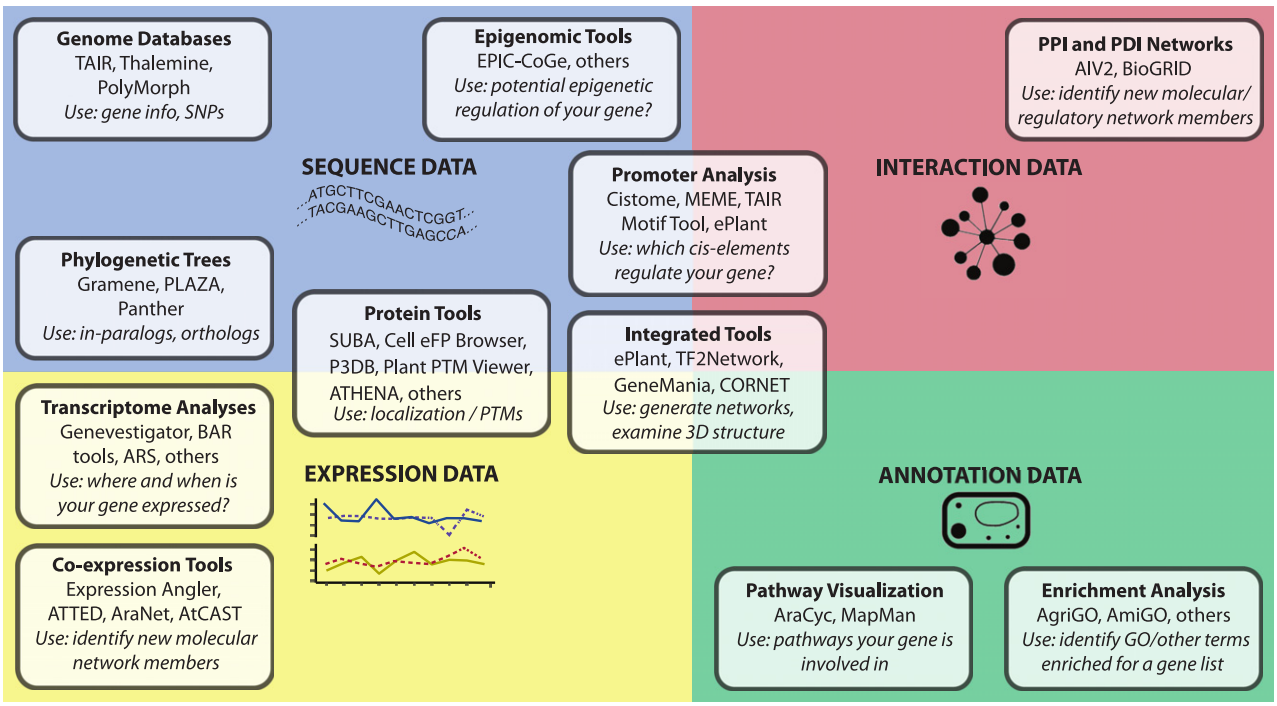

Figure 1. Tools discussed in this review. Their potential uses by Arabidopsis and other plant researchers are listed in each box. Boxes are divided across four broad categories of sequenced-based, interaction-based, expression-based and annotation-based data.

et al., 2013). A detailed protocol describing how to use PANTHER is available (Mi et al., 2019).

Ensembl Plants, Gramene and PLAZA have other, somewhat complementary functions. For example, Ensembl Plants has a nice interface for exploring genome alignments, as well as providing useful variation/single-nucleotide polymorphism (SNP) tracks. With PLAZA it is possible to examine gene-based collinearity and synteny arrangements both within Arabidopsis and between species. PLAZA also offers Workbench that, like the Phylogenetics Module, allows researchers the ability to perform analyses on customized sets of genes. Last, Ensembl Plants and Gramene offer expression information through the European Bioinformatics Institute's Expression Atlas (Papatheodorou et al., 2018), although we'd recommend using other resources instead because of the somewhat limited data sets for Arabidopsis (see the 'Transcript and transcriptome analysis' section.

Epigenomic tools

The EPIC-CoGe Browser (Epigenomes of Plants International Consortium – Comparative Genomics Browser; Nelson et al., 2018) provides a simple way to explore epigenomic data from hundreds of Arabidopsis sequencing experiments. You can upload your own data, add new genomes and share data easily with collaborators, as well as visualize your own data overlaid with other publicly accessible experimental data sets. Data sets of interest can be identified by keyword, such as 'CHH methylation' or 'Arabidopsis'. Check out https://genomevolution.org/wiki/index.php/EPIC-CoGe_Tutorial for easy-to-follow tutorials. With the EPIC-CoGe site it is possible to examine whether there are any potentially relevant epigenomic (chromatin) regulatory marks near your gene of interest or regions of open chromatin, as determined by DNAse I hypersensitivity by the Plant Regulome project (Sullivan et al., 2014). Two other sites that might be of interest are the Jacobsen Lab Epigenomics Browser (https://www.mcdb.ucla.edu/Research/Jacobsen/LabWebSite/P_EpigenomicsData.php), with several published data sets investigating epigenomic changes in a variety of mutants, and the Ecker Laboratory's 1001 Epigenomes Browser (http://neomorph.salk.edu/1001.aj.php; Kawakatsu et al., 2016), where DNA variants in addition to methylated cytosine variants can be compared across hundreds of Arabidopsis accessions. Methylcytosines are characteristic of epigenetic gene silencing (Pikaard and Mittelsten Scheid, 2014), whereas histone acetylation is usually associated with transcriptionally active genes, as are regions of chromatin accessibility and DNAse hypersensitivity. Thus, examining such data sets can give hints about the possible regulation of your favorite gene.

Transcript and transcriptome analyses

Online tools for examining gene transcript abundance can be used instead of, or in addition to, performing quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) or constructing promoter:reporter fusions to

Bioexample 1. Using natural variation in Arabidopsis accessions to explore nucleotide binding site leucine-rich repeat (^{1}NLR) protein interactions.

Hybrid necrosis is a form of incompatibility caused by the activation of immune signaling in the absence of a pathogen trigger (Bomblies and Weigel, 2007). Such an interaction arises when elements of the two different parental immune systems interact in the offspring of a cross between the parents, causing severe cell death and loss of viability. Nucleotide binding site NLR proteins are one of the key elements causing these detrimental interactions (Chae et al., 2014). NLRs are immune receptors that directly or indirectly recognize other molecules, often from pathogens, leading to conformational changes and intramolecular associations that trigger downstream signaling cascades (Jones and Dangl, 2006). Homomeric or heteromeric protein–protein associations are key to the function of several of these NLRs. The association between different natural genetic variants is a primary cause of hybrid necrosis (Chae et al., 2014).

The NLR genes are some of the most polymorphic between genetically varying populations within a species (Cao et al., 2011). In *Arabidopsis thaliana*, over half of the NLRs are located in highly polymorphic multigene clusters (Van de Weyer et al., 2019). With the establishment of low-cost, high-throughput sequencing technologies, the 1001 Genomes Project (https://1001genomes.org,) has sequenced the genomes of 1135 Arabidopsis accessions (1001 Genomes Consortium, 2016). Since its release, the 1001 Genomes Project (see the 'Genome databases' section) has enabled multiple discoveries by exploiting natural genetic variation (e.g. Baudin et al., 2021). Recent advances in our understanding of hybrid necrosis have been accomplished by taking advantage of this and other resources mentioned in this review. In Barragan et al. (2021), once a specific variant was identified via a genome-wide association study (GWAS) of F_1 crosses that trigger hybrid necrosis, this variant was validated using an artificial microRNA (miRNA) designed with the WMD3 online tool (covered in the 'Targeting tools' section). The authors then took advantage of the polymorphism information for accessions of the 1001 Genomes Project to further validate their findings. Using these data, they selectively crossed other accessions carrying their identified variants, which, as predicted, caused hybrid necrosis. Arabidopsis researchers are encouraged to search their genes of interest in the Polymorph tool from the 1001 Genomes Project (covered in the 'Genome databases' section) to check whether the large genetic variation within these sequenced populations can be used to their advantage.

determine where a gene is transcribed. Such inquiries can guide a scientist where to look for a phenotype if no phenotype is apparent under 'standard' conditions or if a phenotype might be apparent in an individual cell type relative to bulk tissue. Web-based tools such as Genevestigator or the eFP Browser make this very easy, with a few caveats: for instance, the older microarray data in these platforms don't provide coverage of the entire gene space in Arabidopsis, but newer RNA sequencing (RNA-seq)-based measurements can help address this issue, and sometimes the number of replicates is low.

The Bio-Analytic Resource for Plant Biology operates the eFP (electronic fluorescent pictograph) Browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi; Winter et al., 2007), which provides easy access to 175 million expression measurements – both RNA-seq and microarray based – from *A. thaliana*, *Glycine max* (soybean), *Hordeum vulgare* (barley), *Medicago truncatula* (barrel medic), *Oryza sativa* (rice), *Populus trichocarpa* (poplar), *Zea mays* (maize) and others. Small pictures depict the plant parts and contexts used to generate the expression data, and a color scale is used to denote expression level. If a given gene is not on the ATH1 array, then RNA-seq atlases (e.g. 'Klepikova Atlas', 'Shoot Apex', 'Embryo', 'Silique' or 'Germination' data sources) maybe be queried.

GENEVESTIGATOR (https://www.genevestigator.com/) provides access to data from more than 10 000 high-quality ATH1 arrays and RNA-seq experiments for Arabidopsis (Hruz et al., 2008). Different tools available within GENEVESTIGATOR let you examine when and where your gene of interest is expressed and in response to which conditions, namely the Anatomy Condition Search tool and the Perturbations Condition Search tool, respectively. In contrast to the eFP Browser, GENEVESTIGATOR queries are returned in a heat-map format, as opposed to pictographically. Thus, it is possible to simultaneously analyze hundreds of genes. This is in contrast to the eFP Browser that allows a user to examine just one gene at a time, although other tools at the BAR permit multi-gene queries. Genevisible (https://genevisible.com/search), a free site operated by GENEVESTIGATOR, which itself requires a subscription or can be trialed for a limited time without payment, can be used to search the developmental or perturbation compendia GENEVESTIGATOR has curated to find the 20 data sets where your favorite gene has the weakest or strongest expression levels.

Another comprehensive database of more than 20 000 RNA-seq data sets from Arabidopsis is the Arabidopsis RNA-Seq database (ARS, http://ipf.sustech.edu.cn/pub/athrna; Zhang et al., 2020). The creators of this database have annotated thousands of RNA-seq samples by experimental treatment, part of the plant, genotype, etc., and made them accessible with an easy-to-use web interface.

Several single-cell RNA-seq databases are worth mentioning: the Root Cell Atlas (http://wanglab.sippe.ac.cn/rootatlas; Zhang et al., 2019), the Plant Single Cell RNA-Sequencing Database (https://www.zmbp-resources.uni-tuebingen.de/timmermans/plant-single-cell-browser; Ma et al., 2020) and a single-cell view within the eFP Browser and ePlant (Waese-Perlman et al., 2021), based on root

data from Ryu et al. (2019). For developmental biologists, these data can help reveal the variability in expression levels for a gene or sets of genes in subpopulations of cells ostensibly from the same region of the plant.

Lastly, the eFP-Seq Browser tool (BAR) permits the visualization of 113 RNA-seq data sets used to create the ARAPORT 11 reannotation of the Arabidopsis genome (Cheng et al., 2017) and a developmental atlas published by Klepikova et al. (2016). The eFP-Seq Browser displays the number of reads mapped above the desired ARAPORT 11 gene model and shows summarized expression levels as small eFP pictographs of the corresponding plant part. The TraVA tool may also be used to explore the data from Klepikova et al. (2016).

Co-expression tools

The 'guilt-by-association' paradigm posits that the genes with similar expression patterns are part of the same biological process as the query gene. Thus, genes without any functional annotation can be identified by this method. The power of co-expression analysis for hypothesis generation is reviewed by Usadel et al. (2009), and more recently by Serin et al. (2016) and Rao and Dixon (2019), although both these reviews puzzlingly omit Expression Angler from their list (see below).

Both the original and the updated Expression Angler (Austin et al., 2016; Toufighi et al., 2005) are user-friendly tools for identifying co-expressed genes, as measured by the Pearson correlation coefficient (*r*), in both a condition-dependent and condition-independent manner. It is often useful to examine condition-dependent data sets, as genes may respond one way in one set of tissues and in the opposite way in other tissues. If one merges these sets, then these correlations cannot be detected (for a discussion of this problem, see Usadel et al., 2009). It is possible to search nine different compendia, and genes with an *r* value of greater than 0.75 can be considered co-expressed. This analysis might help identify closely associated genes that are poorly annotated, and therefore would not be considered interesting candidates at first glance. It is also possible to design a desired pattern of expression with the custom bait feature of Expression Angler. ATTED II is a database for exploring co-expression relationships between genes (Aoki et al., 2016; Obayashi et al., 2018). The latest version of this tool (ver. 11.0) uses the logit score (LS) to report the degree of co-expression between genes, which is a transformation of the mutual rank (MR) of the Pearson's correlation coefficient (Obayashi and Kinoshita, 2009), in a condition-independent way in microarray- or RNA-seq-based expression compendia, or across five compendia of experimental conditions: tissue, biotic stress, abiotic stress, light and hormone treatment. ATTED II also contains pre-computed co-expression data from other species to provide a comparative view across these species using putative gene orthologs. Last, AraNet (Lee et al., 2015) uses machine learning to generate networks of co-functional genes based on a 'gold standard' of reliable Gene Ontology (GO) terms and pathways from AraCyc, and using the guilt-by-association paradigm and likelihood scores.

Another way to use transcriptomic data is to search for correlations between samples. The *Arabidopsis thaliana* Correlation Analysis Tool (AtCAST3) does this across a large number of transcriptomic samples (Kakei and Shimada, 2015), where the expression levels for all genes for a given experiment serve as 'signatures' of that experiment. Using AtCAST3, you can identify other experiments/samples where those genes also exhibit similar signatures of expression. You can do this for all samples in its database, or better yet you can use your own expression data set to identify what transcriptomic experiment most closely resembles your own. You can think of it as BLAST for transcriptomics data, but rather than identifying what other sequences are similar, you are identifying what other experiments are similar, based on their transcriptomic profiles. With AtCAST3, you could quickly form a functional hypothesis as to a role for a mutant gene or mode of action for a chemical.

Promoter/regulatory region analysis

The presence of *cis*-regulatory elements in the gene promoters or distal regulatory regions, in part, controls when and where genes are transcribed. These *cis*-elements can be bound by one or more transcription factors, which regulate transcription. This is an active area of research, and many useful chromatin accessibility, DNA methylation and histone post-translational modification data sets have been published (e.g. Lu et al., 2019; Crisp et al., 2020), but might not have yet made their way into the following resources. We will touch on a few tools here for analyzing *cis*-elements in promoter regions of Arabidopsis genes; however, the Ecker Lab's DAP-seq tool (http://neomorph.salk.edu/aj2/pages/hchen/dap_ath_pub_models.php; O'Malley et al., 2016) and the previously mentioned EPIC-CoGe browser are also useful resources for promoter analysis. EPIC-CoGe can be used to explore epigenetic modifications in promoters, whereas the DAP-seq tool can be used to see whether a given transcription factor from the Ecker Lab's DAP-seq collection was found to bind to a region of interest.

If you have a set of co-expressed genes, either from your own transcriptomic experiment or from one of the co-expression tools mentioned in the previous section, a logical question to ask is whether there are *cis*-elements that are enriched in the promoters of those genes. Cistome (Austin et al., 2016) enables you to search for *cis*-elements that are enriched in the promoters of your genes of interest.

As with Cistome, you can use the MEME Suite (Bailey et al., 2009) to search a set of gene promoters for enriched

cis-elements or to scan promoters of interest to find matches to known transcription factor binding specificities (TFBSs). With a bit of extra work you can scan promoters with a set of TFBSs, including those from the Ecker Lab's Cistrome effort, which documents the transcription factor binding sites for hundreds of Arabidopsis transcription factors (O'Malley et al., 2016). FIMO (Find Individual Motif Occurrences; Grant et al., 2011) and AME (Analysis of Motif Enrichment; McLeay and Bailey, 2010) are of particular interest: with FIMO you can scan one or more promoters for matches to each TFBS in a database (e.g. using supplemental TFBS data from O'Malley et al., 2016); with AME it is possible to search for known TFBSs or motifs that are enriched in the input promoter sequences, as compared with 'background' sequences. Many of the outputs of one MEME Suite tool can be piped (i.e. easily transferred) to another for further analyses. For instance, MEME-ChIP is able to handle thousands of sequences from chromatin immunoprecipitation (ChIP)-seq experiments and provide downstream prediction of novel motifs with the MEME or STREME (Bailey, 2021) tools, with the caveat that the motif–transcription factor association might not be known. Another tool for promoter analyses is ePlant (Waese et al., 2017). With the Interaction Viewer of ePlant you can view interactors of your gene product, which can be informative for promoter analyses if your gene product is a transcription factor. The protein–DNA interaction (PDI) data in ePlant come from DNA Affinity Purification sequencing (DAP-seq) data from O'Malley et al. (2016) and from yeast one-hybrid (Y1H) experiments (Brady et al., 2011; de Lucas et al., 2016; Gaudinier et al., 2011; Li et al., 2014; Murphy et al., 2016; Porco et al., 2016; Sparks et al., 2016; Taylor-Teeples et al., 2015), as well as predicted interactions based on FIMO mapping (Grant et al., 2011). DNA sequences (gene promoters) are displayed as squares and have curved lines indicating interactions with other proteins. The Motif Analysis tool in TAIR can also identify over-represented 6-mer oligos in upstream regions of genes, which can be compared with TFBS databases, e.g. JASPAR (Fornes et al., 2020) or CisBP (Weirauch et al., 2014), to see whether these correspond to a known *cis*-element/TFBS. Alternatively, *in planta* promoter deletion experiments can conclusively demonstrate the requirement of the 6-mer sequence for expression. We direct readers to a recent review article covering these and other tools (Kulkarni and Vandepoele, 2020).

Gene Ontology/functional enrichment analyses

Asking whether certain terms associated with a list of genes are over-represented is one of the 'bread-and-butter' methods in bioinformatics. Such enrichment tests help us to understand and digest the sometimes overwhelming lists of genes from transcriptomic (or other 'omic') experiments. One of the ontology or "term" systems often used for enrichment tests is the Gene Ontology (GO; Ashburner et al., 2000)–a set of categories, described using a controlled vocabulary, into which genes are placed. The top-level categories are biological process (BP), cellular component (CC) and molecular function (MF). Successively more specific categories are found under the main categories. The developers of MAPMAN (Schwacke et al., 2019; Thimm et al., 2004) have generated a similar bin-based functional categorization system, specific for plants. The tools discussed in this section can perform statistical tests to assess whether the number of genes in your gene list associated with a given category is enriched relative to the number expected by chance, thereby helping you to make sense of long lists of genes.

AgriGO (Tian et al., 2017) is an easy-to-use tool for analyzing whether any particular GO terms are enriched in your Arabidopsis gene list. A nice visualization of enriched terms using the same directed acyclic graph structure on which the GO system was developed may be easily generated. A table of enriched GO terms is also provided. AmiGO (Carbon et al., 2009) is not Arabidopsis specific and provides a generic interface for calculating GO term enrichments for any of the species annotated by the GO Consortium. MAPMAN (Schwacke et al., 2019; Thimm et al., 2004) can also be used to perform such enrichment tests, as can the PLAZA 4.5 Workbench (Van Bel et al., 2018). Last, although not a web-based tool, the BiNGO plug-in (Maere et al., 2005) for the popular JAVA-based standalone network analysis program CYTOSCAPE (Christmas et al., 2005) provides a nice list- and network-based representation of enriched GO terms.

Pathway visualization

As suggested in the previous section, one issue with large data sets is obtaining a bird's-eye overview of the results. In the case of metabolic pathways, if you see multiple genes in a given pathway are all increased in transcript abundance after a perturbation, this lends strong support for that pathway being important for the plant's response to that perturbation. A couple of useful pathway visualization tools have been developed that allow you to project quantitative data (e.g. transcriptomic data) onto curated pathway maps.

AraCyC 15.0 (Mueller et al., 2003, but the resource is updated regularly) is the most extensive Arabidopsis pathway database. Curated pathways may be explored within a web-based interface.

A special Omics Viewer tool, which is part of AraCyC, permits overlaying gene expression data (or any other quantitative data keyed by their Arabidopsis gene identifiers) onto pathway maps to generate an overview figure of which pathways are altered, at least at the level of gene expression. Another popular tool for pathway visualization

is MAPMAN (Schwacke et al., 2019; Thimm et al., 2004). This software, and its associated database, has its own classification system called MAPMAN bins for tagging genes and metabolites in terms of metabolic pathways, biological responses, cellular functions, and gene families. In contrast to most of the other tools described in this review, MAPMAN is standalone and must be installed on your own computer. MAPMAN will generate an image of the pathways and genes showing altered regulation, with genes being depicted by colored squares and their associated expression levels encoded in the colors of the squares.

Protein information

Often, it is useful to know information about the gene product of your gene of interest. 'Where is it localized in the cell?' 'Are there any post-translational modifications that might regulate the activity of the protein?' For subcellular localization, the main resource is SUBA4 (Hooper et al., 2017), although TAIR also annotates Arabidopsis gene products with their corresponding GO Cellular Compartment if that has been published. Approximately 40% of Arabidopsis gene products have some sort of experimental data supporting a given localization. Furthermore, the SUBA curators have run 22 subcellular prediction algorithms for all proteins in the Arabidopsis proteome. The SUBA interface provides extensive query options. An alternative way of viewing the data in SUBA is provided by the Cell eFP Browser (Winter et al., 2007, but updated since then!). It generates a pictograph of the predicted and experimentally determined localizations listed in the SUBA database. Cell eFP uses a simple weighting function to give a higher score to 'direct assay' (i.e. literature supported) subcellular localization than to predicted subcellular localizations.

P^3DB is a plant protein phosphorylation (denoted by the P^3) and acetylation database with phosphoproteomic data for many plant species, including *A. thaliana* (Gao et al., 2009; Yao et al., 2012, 2014). Although it has not been recently updated, P^3DB contains curated data describing approximately 50,000 phosphosites and approximately 16,000 phosphoproteins across nine plant species. As with P^3DB, the Plant PTM Viewer is a database of experimentally determined post-translational modifications (PTMs), such as phosphorylation sites, in many plant proteins (437,318 PTMs in 103,975 proteins are stored in the database, of which 165,193 PTMs are found in Arabidopsis proteins). The Plant PTM Viewer contains a wider variety of PTMs, including methylation, nitrosylation, ubiquitination and glycosylation (Willems et al., 2019). You can check out the PTM Viewer tutorial at https://dev.bits.vib.be/ptm-viewer/tutorial.php. More specifically, the Ubiquitination Site tool (http://bioinformatics.psb.ugent.be/webtools/ubiquitin_viewer) provides additional data on protein ubiquitination from Arabidopsis cell culture samples (Walton et al., 2016).

Lastly, you can examine expression profiles and matched quantitative proteomic samples from 30 tissues of Arabidopsis using ATHENA (Mergner et al., 2020). The proteomic samples also include information about phosphorylation.

Protein–protein interaction networks

In a way that is analogous to using co-expressed genes to ascribe function after the 'guilt-by-association' paradigm, querying interaction databases might provide insight into one's gene product of interest by letting you know its interaction partners and their annotations. A fairly comprehensive Arabidopsis-specific interaction database is BAR's Arabidopsis Interactions Viewer 2 (AIV2; Dong et al., 2019), as well as the Arabidopsis-specific AtPID (http://www.megabionet.org/atpid/) by Li et al. (2011). We strongly recommend checking out other databases that are not Arabidopsis-specific, such as BioGRID (http://thebiogrid.org; Chatr-Aryamontri et al., 2017) or IntAct (http://www.ebi.ac.uk/intact/; Orchard et al., 2014). The reason for checking out multiple database is that literature curation efforts are by no means complete for any of these. AIV2 permits you to query several Proteomics Standard Initiative Common Query Interface (PSICQUIC)-enabled databases for other Arabidopsis interactions, thereby making the task of searching of multiple databases easier (Aranda et al., 2011), and includes 2.8 million protein–DNA interactions from the Ecker Lab's DAP-seq data (O'Malley et al., 2016).

Integrated tools

Integrated tools display data from multiple sources for easier visualization and to improve prediction accuracy. The following bioinformatics tools integrate protein and genetic interactions, protein domain similarity, pathways, co-localization, and co-expression, permitting the user to generate hypotheses in a rapid manner. The GeneMANIA (Mostafavi et al., 2008) algorithm uses these different types of data to predict the function for a single gene or to find new members of a protein complex or pathway. ePlant (Waese et al., 2017) is more of a visual analytic tool that combines several common tools for plant biology research. You can examine polymorphisms from the 1001 Genomes Project, visualize gene expression in the whole plant and/or in different tissues, see the subcellular localization of a protein, find its interactors and view predicted or experimentally determined protein structures. A nice feature for the Molecule Viewer is to see where non-synonymous SNPs from the 1001 Genomes Project (1001 Genomes Consortium, 2016) cause amino acid changes, and if those changes might be near to an active site or binding domain. The last integrative tools we recommend are TF2Network (Kulkarni et al., 2018) and CORNET (Van Bel and Coppens, 2017). TF2Network can identify potential

Bioexample 2. A multimodal approach to study promoters and to engineer genetic expression.

Ever since its conception, CRISPR-Cas9 mutagenesis has become one of the most prominent and powerful tools in a geneticist's toolbox. Although CRISPR is primarily used by scientists for generating knockout alleles of their gene of interest, its capabilities have been applied in more 'outside-the-box' approaches. For example, Rodriguez-Leal et al. (2017) applied CRISPR to generate dozens of novel promoter alleles for genes that regulate *Solanum lycopersicum* (tomato) inflorescence and fruit development, such as *CLV3*. First, the authors designed several CRISPR guide RNAs targeting the promoter regions of these genes using the CRISPR-P tool (covered in the 'Targeting tools' section). They then generated transgenic lines carrying a construct with Cas9 and all these gRNAs and crossed these to the wild type. Each cross generated different rearrangements in their respective targeted promoter, resulting in a large number of novel regulatory regions. The new alleles produced using this 'promoter-bashing' approach effectively generated a platform for engineering gene dosage, by which the authors could correlate expression with promoter architecture. They proceeded to identify evolutionarily conserved elements in the *cis*-regulatory regions revealed by the alleles using CoGE. Once these conserved regulatory regions were determined, the authors identified potential transcription factor binding sites (TFBSs) using JASPAR, Cistrome and MEME, leveraging information from Arabidopsis data sets (these types of analyses are covered in the 'Promoter analysis' section of this review). Readers will find these types of approaches are extremely beneficial in cases where researchers are interested in the effects of gene dosage or cases in which 'loss-of-function' alleles of their gene of interest have detrimental effects that make a traditional knockout study impossible.

common regulators of a gene list, such as co-expressed or differentially expressed genes. The 1793 transcription factor position weight matrices (PWMs), including DAP-seq data from O'Malley et al. (2016), are counted and scored using hypergeometric tests. The top 50 most significant TFs are visually reported as predicted regulators. Experimentally determined PPIs and PDIs are also shown in the output, helping you to understand potential gene regulatory networks that might be regulating your genes of interest. In a similar but less user-friendly manner that requires the installation of JAVA, CORNET (which is worth having for the nice CYTOSCAPE displays it generates) will display regulatory and co-expression associations with a few clicks. The transcription factor binding database does not appear to contain DAP-seq data. A review article by Kulkarni and Vandepoele (2020) on inferring gene regulatory networks provides a nice overview of six web tools for regulatory network inference, whereas Ko and Brandizzi (2020) dive deeper into the mechanics of using network-based approaches for understanding gene regulation in plants.

Targeting tools for confirming gene function

Although the above tools might identify new candidates for involvement in a biological process, we would still need to examine their function. The most biologically relevant way to do this is by reverse genetics. We can use three approaches for this purpose: generating deletions using CRISPR-Cas9; silencing target genes using artificial microRNAs (miRNAs); or tapping into the vast collection of T-DNA mutant lines that are available from the stock centers. For the first approach, we need to create CRISPR guide RNAs, which can be achieved with CRISPR-PLANT (Xie et al., 2014) and CRISPR-P (Lei et al., 2014; Liu et al., 2017). In some instances, obtaining or creating loss-of-function mutations for functional analyses is not possible as complete knockouts might be lethal or we may want to titrate the abundance of the transcript in a spatiotemporal manner. Here, artificial miRNAs can be created. The Web MicroRNA Designer 3 (Ossowski et al., 2008) is a web-based app for easily designing these artificial miRNAs. Perhaps one of the most valuable resources available to Arabidopsis researchers are T-DNA mutant collections, developed in the late 1990s and early 2000s (reviewed in O'Malley and Ecker, 2010). Identifying a potential knockout line on the SiGNAL website (Alonso et al., 2003) is straightforward. The companion website (http://signal.salk.edu/tdnaprimers.2.html) can be used to generate genotyping primers.

Outstanding questions and challenges

- How should we visualize and use data from the Plant Cell Atlas (Rhee et al., 2019) to further understand plants?
- There are continuing issues with site maintenance (AtCAST and AraNet both use a FLASH-based player for output, which is now deprecated, and it is unclear who will modify the code), curation (who adds new data to existing database and how often?) and funding (long-term funding for databases is seldom available).
- Updated genome versions will confound interpretations of derived data, such as expression level estimates based on RNA-seq data.
- How can data be integrated across scales, species and environments?
- Existing TF binding site databases do not profile all TFs in the genome and an absence of a binding site entry may lead to false negatives when querying those databases.

CONCLUSION

In summary, it is clear that there is an enormous volume of data available to Arabidopsis researchers, and that by accessing these data with the tools described in this

review, and others, insights can be garnered that will help with elucidating the role of a gene or set of genes in question. Keep an eye out for new data sets and tools: a new Arabidopsis Lipid Map has just been published (Kehelpannala et al., 2021). As in many areas of human endeavor, change is the only constant!

DATA AVAILABILITY

All relevant data (and URLs for the tools described in this review) can be found within the article and its supporting materials.

ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for helpful comments on an initial draft of this article. NJP was funded by Genome Canada/ Ontario Genomics (OGI-128) to develop the ePlant tool described in this review.

AUTHOR CONTRIBUTIONS

All authors contributed sections to this article, with approximately equal word counts. NJP merged these sections and edited the document. AC-P created Figure 1 and provided the two bioexamples.

CONFLICT OF INTEREST

NJP runs the Bio-Analytic Resource for Plant Biology, where several of the tools described in this review may be found. The authors otherwise declare no conflicts of interest associated with this work.

REFERENCES

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P. *et al.* (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653–657.

Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics (Oxford, England)*, **15**, 391–412.

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K. & Obayashi, T. (2016) ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. *Plant and Cell Physiology*, **57**, e5.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S.L., Ceol, A., Chautard, E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods*, **8**, 528–529.

Ashburner, M., Ball, C.A., Blake, J.A. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Austin, R.S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T.T. *et al.* (2016) New BAR tools for mining expression data and exploring Cis-elements in Arabidopsis thaliana. *The Plant Journal*, **88**, 490–504.

Bailey, T.L. (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834–2840.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, **37**, W202–208.

Barragan, A.C., Collenberg, M., Wang, J., Lee, R.R.Q., Cher, W.Y., Rabanal, F.A. *et al.* (2021) A truncated singleton NLR causes hybrid necrosis in Arabidopsis thaliana. *Molecular Biology and Evolution*, **38**, 557–574.

Baudin, M., Martin, E.C., Sass, C. *et al.* (2021) A natural diversity screen in Arabidopsis thaliana reveals determinants for HopZ1a recognition in the ZAR1-ZED1 immune complex. *Plant, Cell and Environment*, **44**, 629–644.

Bomblies, K. & Weigel, D. (2007) Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics*, **8**, 382–393.

Brady, S.M. & Provart, N.J. (2009) Web-queryable large-scale data sets for hypothesis generation in plant biology. *The Plant Cell*, **21**, 1034–1051.

Brady, S.M., Zhang, L., Megraw, M. *et al.* (2011) A stele-enriched gene regulatory network in the Arabidopsis root. *Molecular Systems Biology*, **7**, 459.

Cao, J., Schneeberger, K., Ossowski, S. *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, **43**, 956–963.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B. & Lewis, S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

Chae, E., Bomblies, K., Kim, S.-T., Karelina, D., Zaidem, M., Ossowski, S. *et al.* (2014) Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell*, **159**, 1341–1351.

Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, **45**, D369–D379.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S. & Town, C.D. (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal*, **89**, 789–804.

Chory, J., Ecker, J.R., Briggs, S., Caboche, M., Coruzzi, G.M., Cook, D. *et al.* (2000) National Science Foundation-Sponsored Workshop Report: "The 2010 Project" Functional Genomics and the Virtual Plant. A Blueprint for Understanding How Plants Are Built and How to Improve Them. *Plant Physiology*, **123**, 423–426.

Christmas, R., Avila-Campillo, I., Bolouri, H. *et al.* (2005) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *American Association of Cancer Research Education Book*, 12–16.

Crisp, P.A., Marand, A.P., Noshay, J.M., Zhou, P., Lu, Z., Schmitz, R.J. & Springer, N.M. (2020) Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. Proc. Natl. Acad. Sci. 117 (38) 23991-24000.

de Lucas, M., Pu, L.I., Turco, G., Gaudinier, A., Morao, A.K., Harashima, H. *et al.* (2016) Transcriptional regulation of Arabidopsis polycomb repressive complex 2 coordinates cell-type proliferation and differentiation. *The Plant Cell*, **28**, 2616–2631.

Dong, S., Lau, V., Song, R. *et al.* (2019) Proteome-wide, structure-based prediction of protein-protein interactions/new molecular interactions viewer. *Plant Physiology*, **179**, 1893–1907.

Fornes, O., Castro-Mondragon, J.A., Khan, A. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **48**, D87–D92.

Gao, J., Agrawal, G.K., Thelen, J.J. & Xu, D. (2009) P3DB: a plant protein phosphorylation database. *Nucleic Acids Research*, **37**, D960–962.

Gaudinier, A., Zhang, L., Reece-Hoyes, J.S. *et al.* (2011) Enhanced Y1H assays for Arabidopsis. *Nature Methods*, **8**, 1053–1055.

Genomes Consortium (2016) 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell*, **166**, 481–491.

Grant, C.E., Bailey, T.L. & Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Hooper, C.M., Castleden, I.R., Tanz, S.K., Aryamanesh, N. & Millar, A.H. (2017) SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Research*, **45**, D1064–D1074.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L. *et al.* (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics*, **2008**, 420747.

Jones, J.D.G. & Dangl, J.L. (2006) The plant immune system. *Nature*, **444**, 323–329.

Kakei, Y. & Shimada, Y. (2015) AtCAST3.0 update: a web-based tool for analysis of transcriptome data by searching similarities in gene expression profiles. *Plant and Cell Physiology*, **56**, e7.

Kawakatsu, T., Huang, S.-S., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A. *et al.* (2016) Epigenomic diversity in a global collection of Arabidopsis thaliana accessions. *Cell*, **166**, 492–505.

Kehelpannala, C., Rupasinghe, T., Pasha, A., Esteban, E., Hennessy, T., Bradley, D. *et al.* (2021) An Arabidopsis lipid map reveals differences between tissues and dynamic changes throughout development. The Plant Journal, Available at: https://doi.org/10.1111/tpj.15278 [Accessed May 13, 2021].

Kersey, P.J., Allen, J.E., Allot, A., Barba, M., Boddu, S., Bolt, B.J. *et al.* (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, **46**, D802–D808.

Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S., Logacheva, M.D. & Penin, A.A. (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *The Plant Journal*, **88**, 1058–1070.

Ko, D.K. & Brandizzi, F. (2020) Network-based approaches for understanding gene regulation and function in plants. *The Plant Journal*, **104**, 302–317.

Krishnakumar, V., Hanlon, M.R., Contrino, S., Ferlanti, E.S., Karamycheva, S., Kim, M. *et al.* (2015) Araport: the Arabidopsis information portal. *Nucleic Acids Research*, **43**, D1003–D1009.

Kulkarni, S.R. & Vandepoele, K. (2020) Inference of plant gene regulatory networks using data-driven methods: a practical overview. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, **1863**, 194447.

Kulkarni, S.R., Vaneechoutte, D., Van de Velde, J. & Vandepoele, K. (2018) TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Research*, **46**, e31.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M. & Rhee, S.Y. (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature Biotechnology*, **28**, 149–156.

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J. *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 Research*, **43**, D996–D1002.

Lei, Y., Lu, L., Liu, H.-Y., Li, S., Xing, F. & Chen, L.-L. (2014) CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Molecular Plant*, **7**, 1494–1496.

Li, B., Gaudinier, A., Tang, M. *et al.* (2014) Promoter-based integration in plant defense regulation. *Plant Physiology*, **166**, 1803–1820.

Li, P., Zang, W., Li, Y., Xu, F., Wang, J. & Shi, T. (2011) AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for Arabidopsis. *Nucleic Acids Research*, **39**, D1130–D1133.

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K. & Chen, L.-L. (2017) CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Molecular Plant*, **10**, 530–532.

Lu, Z., Marand, A.P., Ricci, W.A., Ethridge, C.L., Zhang, X. & Schmitz, R.J. (2019) The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, **5**, 1250–1259.

Ma, X., Denyer, T. & Timmermans, M.C.P. (2020) PscB: a browser to explore plant single cell RNA-sequencing data sets. *Plant Physiology*, **183**, 464–467.

Maere, S., Heymans, K. & Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, **21**, 3448–3449.

Mason, G.A., Cantó-Pastor, A., Brady, S.M. & Provart, N.J. (2021) Bioinformatic Tools in Arabidopsis Research. In J. J. Sanchez-Serrano and J. Salinas, eds. *Arabidopsis Protocols*. Methods in Molecular Biology. New York, NY: Springer US, pp. 25–89. https://doi.org/10.1007/978-1-0716-0880-7_2 [Accessed May 13, 2021].

McLeay, R.C. & Bailey, T.L. (2010) Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A. *et al.* (2020) Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature*, **579**, 409–414.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. & Thomas, P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, **38**, D204–D210.

Mi, H., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, **8**, 1551–1566.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X. *et al.* (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols*, **14**, 703.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, **9**, S4.

Mueller, L.A., Zhang, P. & Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology*, **132**, 453–460.

Murphy, E., Vu, L.D., Van den Broeck, L. *et al.* (2016) RALFL34 regulates formative cell divisions in Arabidopsis pericycle during lateral root initiation. *Journal of Experimental Botany*, **67**, 4863–4875.

Nelson, A.D.L., Haug-Baltzell, A.K., Davey, S., Gregory, B.D. & Lyons, E. (2018) EPIC-CoGe: managing and analyzing genomic data. *Bioinformatics*, **34**, 2651–2653.

O'Malley, R.C. & Ecker, J.R. (2010) Linking genotype to phenotype using the Arabidopsis unimutant collection. *The Plant Journal Cell Molecular Biology*, **61**, 928–940.

O'Malley, R.C., Huang, S.-S.-C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R. *et al.* (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y. & Kinoshita, K. (2018) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant and Cell Physiology*, **59**, e3.

Obayashi, T. & Kinoshita, K. (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Research*, **16**, 249–260.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, **42**, D358–363.

Ossowski, S., Schwab, R. & Weigel, D. (2008) Gene silencing in plants using artificial microRNAs and other small RNAs. *The Plant Journal Cell Molecular Biology*, **53**, 674–690.

Papatheodorou, I., Fonseca, N.A., Keays, M., Tang, Y., Barrera, E., Bazant, W. *et al.* (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, **46**, D246–D251.

Pasha, A., Subramaniam, S., Cleary, A., Chen, X., Berardini, T.Z. & Farmer, A. *et al.* (2020) Araport Lives: An Updated Framework for Arabidopsis Bioinformatics. *The Plant Cell*, 32, 2683–2686.

Pikaard, C.S. & Mittelsten Scheid, O. (2014) Epigenetic regulation in plants. *Cold Spring Harbor Perspectives in Biology*, **6**, a019315.

Porco, S., Larrieu, A., Du, Y. *et al.* (2016) Lateral root emergence in Arabidopsis is dependent on transcription factor LBD29 regulation of auxin influx carrier LAX3. *Development (Cambridge, England)*, **143**, 3340–3349.

Rao, X. & Dixon, R.A. (2019) Co-expression networks for plant biology: why and how. *Acta Biochimica Et Biophysica Sinica*, **51**, 981–988.

Reiser, L., Subramaniam, S., Li, D. & Huala, E. (2017) Using the Arabidopsis information resource (TAIR) to find information about Arabidopsis genes. *Current Protocols in Bioinformatics*, **60**, 1.11.1-1.11.45.

Reuter, J.A., Spacek, D.V. & Snyder, M.P. (2015) High-throughput sequencing technologies. *Molecular Cell*, **58**, 586–597.

Rhee, S.Y., Birnbaum, K.D. & Ehrhardt, D.W. (2019) Towards building a plant cell atlas. *Trends in Plant Science*, **24**, 303–310.

Rodríguez-Leal, D., Lemmon, Z.H., Man, J., Bartlett, M.E. & Lippman, Z.B. (2017) Engineering quantitative trait variation for crop improvement by genome editing. *Cell*, **171**, 470–480.e8.

Ryu, K.H., Huang, L., Kang, H.M. & Schiefelbein, J. (2019) Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiology*, **179**, 1444.

Schwacke, R., Ponce-Soto, G.Y., Krause, K. *et al.* (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, **12**, 879–892.

Serin, E.A.R., Nijveen, H., Hilhorst, H.W.M. & Ligterink, W. (2016) Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science*, **7**, 444.

Sparks, E.E., Drapek, C., Gaudinier, A. *et al.* (2016) Establishment of expression in the SHORTROOT-SCARECROW transcriptional cascade through opposing activities of both activators and repressors. *Developmental Cell*, **39**, 585–596.

Sullivan, A.M., Arsovski, A.A. & Lempe, J. *et al.* (2014) Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Reports, 8, 2015-2030.

Sullivan, A., Purohit, P.K., Freese, N.H. *et al.* (2019) An 'eFP-Seq Browser' for visualizing and exploring RNA sequencing data. *The Plant Journal*, **100**, 641–654.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A. *et al.* (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature*, **517**, 571–575.

Thimm, O., Bläsing, O., Gibon, Y. *et al.* (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal Cell Molecular Biology*, **37**, 914–939.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z. *et al.* (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, **45**, W122–W129.

Togninalli, M., Seren, Ü., Freudenthal, J.A. *et al.* (2020) AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. *Nucleic Acids Research*, **48**, D1063–D1068.

Toufighi, K., Brady, S.M., Austin, R., Ly, E. & Provart, N.J. (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *The Plant Journal*, **43**, 153–163.

Usadel, B., Obayashi, T., Mutwil, M. *et al.* (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell and Environment*, **32**, 1633–1651.

Van Bel, M. & Coppens, F. (2017) Exploring plant co-expression and gene-gene interactions with CORNET 3.0. *Methods Mol. Biol. Clifton NJ*, **1533**, 201–212.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y. *et al.* (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research*, **46**, D1190–D1196.

Van de Weyer, A.-L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K. *et al.* (2019) A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. *Cell*, **178**, 1260–1272.e14.

Waese, J., Fan, J., Pasha, A., Yu, H., Fucile, G., Shi, R. *et al.* (2017) ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *The Plant Cell*, **29**, 1806–1821.

Waese-Perlman, B., Pasha, A., Ho, C., Azhieh, A., Liu, Y., Sullivan, A. *et al.* (2021) ePlant in 2021: New Species, Viewers, Data Sets, and Widgets. bioRxiv, 2021.04.28.441805.

Walton, A., Stes, E., Cybulski, N. *et al.* (2016) It's time for some "Site"-Seeing: novel tools to monitor the ubiquitin landscape in Arabidopsis thaliana. *The Plant Cell*, **28**, 6–16.

Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, **38**, W214–220.

Weirauch, M., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

Willems, P., Horne, A., Parys, T.V., Goormachtig, S., Smet, I.D., Botzki, A., Breusegem, F.V. & Gevaert, K. (2019) The Plant PTM Viewer, a central resource for exploring plant protein modifications. *The Plant Journal*, **99**, 752–762.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. & Provart, N.J. (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One*, **2**, e718.

Xie, K., Zhang, J. & Yang, Y. (2014) Genome-wide prediction of highly specific guide RNA spacers for CRISPR–Cas9-mediated genome editing in model plants and major crops. *Molecular Plant*, **7**, 923–926.

Yao, Q., Bollinger, C., Gao, J., Xu, D. & Thelen, J.J. (2012) P(3)DB: an integrated database for plant protein phosphorylation. *Frontiers in Plant Science*, **3**, 206.

Yao, Q., Ge, H., Wu, S., Zhang, N., Chen, W., Xu, C. *et al.* (2014) P^3DB 3.0: from plant phosphorylation sites to protein networks. *Nucleic Acids Research*, **42**, D1206–1213.

Zhang, H., Zhang, F., Yu, Y., Feng, L., Jia, J., Liu, B. *et al.* (2020) A Comprehensive online database for exploring ~20,000 public Arabidopsis RNA-Seq libraries. *Molecular Plant*, **13**, 1231–1233.

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D. & Wang, J.-W. (2019) A single-cell RNA sequencing profiles the developmental landscape of Arabidopsis root. *Molecular Plant*, **12**, 648–660.