Bioinformatic Tools in *Arabidopsis* Research

G. Alex Mason[1a], Alex Cantó-Pastor[1a], Siobhan M. Brady[a], Nicholas J. Provart[b]

1. Co-first authors
a. Department of Plant Biology, University of California Davis
b. Department of Cell and Systems Biology, University of Toronto
email. nicholas.provart@utoronto.ca

This chapter is a revision of a chapter with the same name by Miguel de Lucas, Nicholas J. Provart and Siobhan Brady in *Arabidopsis Protocols* (2014, 1062, p. 97-136), edited by José Juan Sanchez Serrano. All material has been revised and updated as of May 2019, and several new tools are described.

Background

Bioinformatic tools are now an everyday part of a plant researcher's collection of protocols. They allow almost instantaneous access to large data sets encompassing genomes, transcriptomes, proteomes, epigenomes and other "-omes", which are now being generated with increasing speed and decreasing cost. With the appropriate queries, such tools can generate quality hypotheses, sometimes without the need for new experimental data. In this chapter, we will investigate some of the tools used for examining gene expression and coexpression patterns, performing promoter analyses and functional classification enrichment for sets of genes, and exploring protein-protein and protein-DNA interactions. We will also cover additional tools that allow integration of data from several sources for improved hypothesis generation.

Keywords: transcriptomics, hypothesis generation, eFP, bioinformatics, *in silico*, proteomics, protein-protein interactions, coexpression, functional classification, functional genomics, promoter analysis, subcellular localization.

1 Introduction

The past decade has been transformative for plant biology. Numerous sequencing-based methods have enabled high-throughput analysis of genomes, epigenomes, transcriptomes, and protein-protein or protein-DNA interactions (Reuter et al., 2015). Other high-throughput methods have enabled quantitative proteome and metabolome measurements. While each individual data set has been of obvious value to the plant biologist who created it, once publicly-available these data sets are useful to plant biologists around the world for querying in the context of their own biological questions (Chory et al., 2000). Such large data sets may not provide a complete understanding of a given system, but they can be leveraged to help plan experiments or generate hypotheses *in silico*. These hypotheses can then be rapidly tested in the lab with the wide range of molecular techniques and genetic resources currently available. This chapter aims to provide an updated overview of web-based tools for querying data sets generated by researchers, often funded by the National Science Foundation Arabidopsis 2010 project in the U.S., whose stated goal was to identify the functions of 25,000 genes in *Arabidopsis* by 2010 (Chory et al., 2000), and by the AtGenExpress Consortium, an international effort to measure the *Arabidopsis* transcriptome under many conditions and in different tissues.

Here, we will emphasize well-cited web-based tools that integrate data from several repositories – tools that draw from many sources are often more useful to the typical *Arabidopsis* researcher than single data source lab-based websites. Additionally, we would like to refer you to a new unit on The Arabidopsis Information Resource (TAIR at http://www.arabidopsis.org) published recently by Eva Huala and colleagues that covers this excellent sequence-centric Arabidopsis database (Reiser et al., 2017). The SIGnAL website at http://signal.salk.edu/ (Alonso et al., 2003) and https://www.araport.org (Krishnakumar et al., 2015) are two further websites for exploring sequences and identifying insertions – we will touch on these briefly, along with websites for the 1001 Arabidopsis genomes project.

The main focus of the chapter will be on tools for exploring transcriptome data sets, which are the most comprehensive of all the large data types, and highlight the ones used for querying these data sets both in targeted and correlative ways. Such tools can be highly valuable for focusing the search for mutant phenotypes, or for providing leads on novel genetic associations with a given biological process, respectively. We will also examine several resources for exploring protein-protein interactions in *Arabidopsis*, or for performing promoter analyses. Integrating different data types to improve function prediction is key to extracting even more knowledge from these data sets.

As in the first edition of this chapter, we will use *ABSCISIC ACID INSENSITIVE 3*, *At3g24650* (Finkelstein and Somerville, 1990) as our “gene of interest”. While this gene has long been known to be involved in seed biology, we will hypothesize some additional functions using the tools described here, some easily inferred at the cost of only a click of the mouse. The programs and websites that will be discussed in this chapter are listed in **Table 1** in the Materials section. Two further informative, if slightly older, review articles in the context of bioinformatic tools for hypothesis generation are by Brady and Provart (Brady and Provart, 2009) and by Usadel and colleagues (Usadel et al., 2009). We would also like to point you to a recent article on the future of Arabidopsis informatics resources (IAIC, 2019). There, the International Arabidopsis Informatics Consortium worked to create a “super-portal” to keep track of and functionally “tag” various Arabidopsis tools - see https://conf.arabidopsis.org/display/COM/Resources. Finally, one of the co-authors on this book chapter has released a free course on Coursera.org called Plant Bioinformatics at https://www.coursera.org/learn/plant-bioinformatics/. Many of the tools described in this chapter are covered in this online lab course.

2 Materials

Table 1. Tools, URLs, and References.

Methods	Tool	Web	Ref.
3.1 Genome Browsers	Araport	https://www.araport.org/	(Krishnakumar et al., 2015)
	1001 Genomes	https://1001genomes.org/	(1001 Genomes Consortium, 2016)
3.2 Precomputed Gene Trees	Ensembl Plants Compara	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)
3.3 Epigenomic Tools	EPIC-CoGe	https://genomevolution.org/CoGe/User.pl	(Nelson et al., 2018)
3.4 Expression Analysis	eFP Browser / eFP-Seq Browser	http://bar.utoronto.ca; http://bar.utoronto.ca/eFP-Seq_Browser/	(Winter et al., 2007)
	Genevestigator	www.genevestigator.com/gv/; https://genevisible.com/search	(Hruz et al., 2008)
	TravaDB, NCBI	http://travadb.org	(Klepikova et al., 2016)
3.5 Coexpression 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)
3.6 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)
3.7 Functional Classification	AgriGO	http://systemsbiology.cau.edu.cn/agriGOv2/	(Tian et al., 2017)
	AmiGO	http://amigo.geneontology.org/rte	(Carbon et al., 2009)
	Classification SuperViewer	http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi	(Provart and Zhu, 2003)
3.8 Pathway Visualization	AraCyc	www.plantcyc.org/	(Mueller et al., 2003)
	MapMan	http://mapman.gabipd.org/web/guest/mapman-download	(Thimm et al., 2004)
3.9 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^{3}DB – Plant Protein Phosphorylation DB	http://p3db.org/index.php	(Gao et al., 2009; Yao et al., 2014, 2012)
	Plant PTM Viewer	https://dev.bits.vib.be/ptm-viewer/index.php	(Willems et al., 2019)
3.10 Protein-Protein Interaction	Arabidopsis Interactions Viewer 2	http://bar.utoronto.ca/interactions2/	(Dong et al., 2019)
3.11 Integrated Tools	Virtual Plant	http://virtualplant.bio.nyu.edu/cgi-bin/vpweb/	(Katari et al., 2010)
	Gene Mania	http://genemania.org/	(Warde-Farley et al., 2010)
	ePlant	http://bar.utoronto.ca	(Waese et al., 2017)
3.12 Targeting Tools	CRISPR-PLANT	https://www.genome.arizona.edu/crispr/CRISPRsearch.html	(Xie et al., 2014)
	WMD3	http://wmd3.weigelworld.org	(Ossowski et al., 2008)
	SIGnAL T-DNA Express	http://signal.salk.edu/	(Alonso et al., 2003)

Supplementary Table 1: *ABI3* developmentally coexpressed genes

AT4G27160
AT2G33520
AT1G14950
AT4G27140
AT2G23640
AT2G02120
AT4G26740
AT5G24950
AT4G27460
AT5G55240
AT5G54740
AT1G17810
AT1G48130
AT5G50360
AT1G65090
AT2G27380
AT4G27150
AT3G44830
AT1G05510
AT5G01300
AT5G01670
AT3G18570
AT2G02580
AT1G73190
AT1G80090
AT3G22640
AT3G54940
AT1G54860
AT2G34315
AT1G52690
AT3G14360
AT3G24650
AT1G03890
AT5G50600
AT5G10140
AT2G41070
AT5G57390
AT1G27461
AT5G60460
AT4G16160
AT3G62730
AT4G10020
AT5G24130
AT1G04560
AT2G21490
AT1G62710
AT2G28490
AT4G31830
AT1G32560
AT2G38905
AT1G29680

Additionally, we use a list of genes differentially regulated in a LEC1 overexpressor as outlined in Mu et al. (2008)[1].

3 Methods

3.1 Genome Databases

As mentioned in the introduction, we would like to refer you to a recent protocol by the curators at The Arabidopsis Information Resource (TAIR, www.arabidopsis.org) on how to access the information available at TAIR, one of the most widely used Arabidopsis portals (Reiser et al., 2017). We try to use *ABI3* to refer to the gene and ABI3 to refer to the protein, but when describing text to be entered into web pages, we don't make the distinction. We also use Arabidopsis to refer to *Arabidopsis thaliana*.

3.1.1 Araport

Another good place to start for information about *Arabidopsis* genes is Araport. Short for *Arabidopsis* Information Portal, it aggregates data from published literature and data sets, and provides computational and visualization tools to help with *in silico* analyses. With the following steps we are able to access both the DNA sequence and the protein sequence of the *ABI3* gene and ABI3 protein.

1. Go to https://www.araport.org/ and locate the search box, surrounded by green[2].

2. Type "ABI3" into the box and click on the magnifying glass.

3. On the resulting page, click on the first result, which should be labelled AT3G24650, of type "Gene".

[1] Available at http://bar.utoronto.ca/~nprovart/Genes_up_in_Lec1OX_w_decimal.txt
[2] Not the case in Safari!

4. The link takes us to ThaleMine, which is a module of Araport.

5. A summary of the information about the *ABI3* gene product is presented at the top of the page.

6. Click on “Genomics” on the “Quick Links” bar. You can obtain the DNA sequence of *ABI3* by clicking on the “FASTA” button here.

7. If you click on “Proteins” and click on the first label (ABI3_ARATH) under “DB Identifier” you will get more information about the protein. A similar “FASTA” button exists to download the ABI protein’s amino acid sequence.

3.1.2 1001 Genomes

The 1001 Genomes project provides data on more than 1001 strains (accessions or genotypes) of *Arabidopsis thaliana* collected from around the world (1001 Genomes Consortium, 2016). A single reference genome is now thought to not be sufficient to represent the genes of an entire species, considering how many polymorphisms have been found to occur across the vast number of sequenced accessions. With the PolyMorph tool we can examine if there are any synonymous or non-synonymous variants, or small insertions or deletions in naturally occurring alleles of a gene of interest based on the sequence data that have been generated by the 1001 Genomes project.

1. Go to http://tools.1001genomes.org/polymorph/. Under the “Locus” parameter, input “AT3G24650” as the “Gene Identifier”. Select “AT3G24650.1”.

2. Under “Filter”, set the “Variant Type” to “SNPs”, “Impact” to “All”, and “Effect Type” to “All”. Click on the “Accessions” parameter to see some of the accessions, but do not deselect any accessions. You can go to http://1001genomes.org/accessions.html to learn more about them. Click Search.

3. You will see a table of results, sorted by position, although you can change this by clicking on the table headers (this will take a moment). Click on Strain to sort by strain number. The first sequenced *Arabidopsis* strain was Col-0. In 1001 Genomes its strain number is 6909. Note that polymorphisms for Col-0 will not be present in the table as the polymorphisms are presented relative to this strain!

4. Look at the “Effect Type” column and click on the table rows to see details. In particular, look for an effect where the impact is “MODERATE”. These kinds of mutations are typically non-synonymous nucleotide changes that result in a changed amino acid in one or more strains *of Arabidopsis thaliana*. These may subtly change the properties of the protein. For instance, the *A. thaliana* ecotype Ale-Stenar-44-4 (strain 992) seems to have non-synonymous single nucleotide polymorphism in *ABI3* that causes a proline to leucine change at amino acid 556 of ABI3, which is near the DNA binding site (based on

information we could see at Araport), a change that might cause ABI3 to bind differently to DNA in that ecotype.

5. You can also look for small insertions in *ABI3* other strains. Switch the “Variant Type” to “Insertions”. Keep “Impact” and “Effect Type” at “All”.

6. To search for small deletions, switch the “Variant Type” to “Deletions”. Keep “Impact” and “Effect Type” at “All”.

3.2 Precomputed Gene Trees

The recent explosion in the number of available sequenced genomes, a resource once reserved only for a handful of model species, has boosted the value of phylogenomic approaches in understanding gene function. Investigating the evolutionary history of a gene can provide valuable insight into its potential role. As an example, imagine we had identified a mutation of a gene, but its annotation revealed little about its biological function. We could investigate its evolutionary relationships and use the annotations of its homologs to help guide us towards a function. Alternatively, this approach would also reveal additional relevant information such as gene duplication events that might help us refine our subsequent reverse genetic strategies (*e.g.* generating a double knock-out mutant) or generate hypotheses about sub-functionalization events. Finally, if we wished to determine the function of a given Arabidopsis gene in another species, we could identify the homolog most likely to be the ortholog. For these purposes, we can take advantage of precomputed phylogenetic trees, available from online platforms such as Ensembl Plants or PLAZA.

3.2.1 Ensembl Plants

Ensembl Plants is a scientific project aimed to provide genome-scale data from plant species, as part of the Ensembl Genomes initiative (Kersey et al., 2018). Ensembl Plants was created as joint effort between the European Bioinformatics Institute and the teams at Gramene. It is maintained by the EBI and is based on the Ensembl annotation framework (Hubbard et al., 2002).

1. Go to https://plants.ensembl.org/index.html. In the search box, search for “ABI3” and select “*Arabidopsis thaliana*” as the species. Click on the first result (labeled “ABI3”).

2. Find the display menu, located on the left sidebar. Also note that Ensembl Plants has many built-in search tools, like BLAST and HMMER, linked in the top navigation bar.

3. Click on the “Gene tree” link on the left sidebar (under “Plant Compara”). You will see some information at the top, and then a graphic similar to **Figure 1**, but interactive.

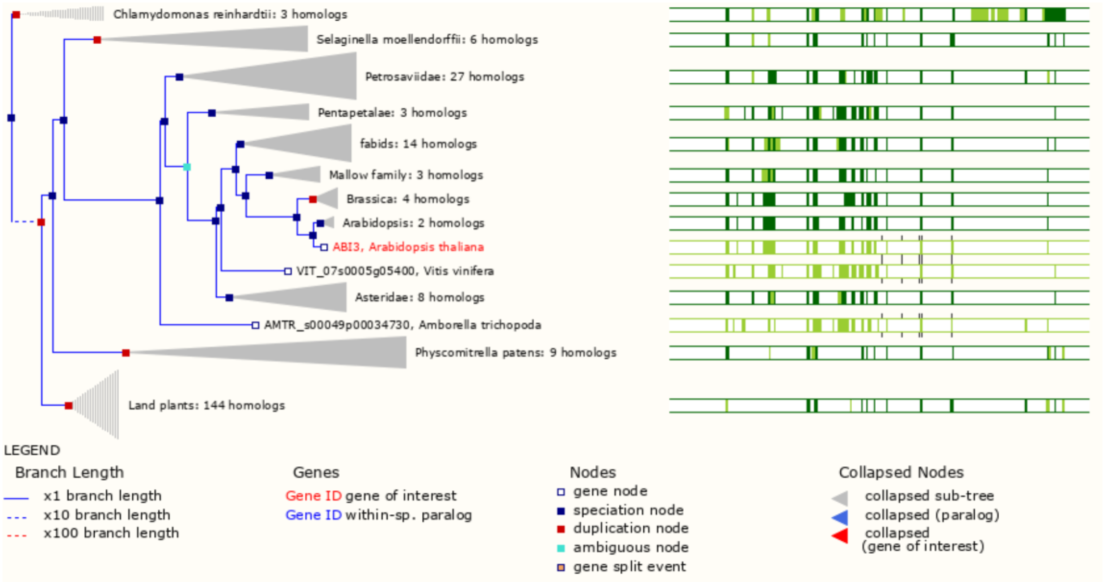

Chlamydomonas reinhardtii: 3 homologs
Selaginella moellendorffii: 6 homologs
Petrosaviidae: 27 homologs
Pentapetalae: 3 homologs
fabids: 14 homologs
Mallow family: 3 homologs
Brassica: 4 homologs
Arabidopsis: 2 homologs
ABI3, Arabidopsis thaliana
VIT_07s0005g05400, Vitis vinifera
Asteridae: 8 homologs
AMTR_s00049p00034730, Amborella trichopoda
Physcomitrella patens: 9 homologs
Land plants: 144 homologs
LEGEND
Branch Length
x1 branch length
x10 branch length
x100 branch length
Genes
Gene ID gene of interest
Gene ID within-sp. paralog
Nodes
gene node
speciation node
duplication node
ambiguous node
gene split event
Collapsed Nodes
collapsed sub-tree
collapsed (paralog)
collapsed (gene of interest)

Figure 1: Ensembl Plants gene tree for *ABI3*.

4. You can collapse and expand branches in the tree by clicking on the squares (nodes) and triangles (subtrees). Try clicking on the *Petrosaviidae* (monocot clade) subtree. A menu will appear, giving the option to expand the sub-tree, amongst other options.

5. Exploring this tree allows us to identify previously characterized homologs in other species, such as *Viviparous-1* (*Vp1*) in *Zea mays*. This homolog has also long been known to be a transcription factor involved in seed maturation and ABA response (McCarty et al., 1989).

6. Additionally, above the graphic there are radio buttons corresponding to Gene Ontology terms or InterPro domains. Selecting a GO term will highlight the genes that are tagged with it in green.

With the phylogenetic trees available at Ensembl Plants, in addition to being able to easily identify homologs in other species, we can observe some interesting patterns. The striking presence of 9 homologs of *ABI3* in *Physcomitrella patens*, a moss, almost begs further investigation.

3.2.2 PLAZA

PLAZA is an integrated resource similar to TAIR and Araport, and in fact crosslinks to these databases and often uses the same sources. It aims to centralize genomic data produced by different genome sequencing initiatives to provide comparative genomic results (Van Bel et al., 2018).

1. Go to https://bioinformatics.psb.ugent.be/plaza/. Here you can find all of the species supported by PLAZA. Select the appropriate PLAZA instance (in this case “Dicots PLAZA 4.0”).

2. Search for “AT3G24650” under Dicots PLAZA (choose “Search” in the menu bar).

3. Scroll down to the Toolbox. Click on the option to view the phylogenetic tree of the homologous gene family. This will display a phylogenetic tree with the information of all plant species stored in PLAZA.

4. Since the previous tree can be difficult to navigate due to the overwhelming amount of information, PLAZA V4 has a function to create a custom tree using the Interactive Phylogenetics Module in its toolbox. **Figure 2** is one such tree, made by selecting the members in *Arabidopsis thaliana* for HOM04D000569, the family in which *ABI3* is located, under Data Settings and Species Selection, and using default settings under Program Settings.

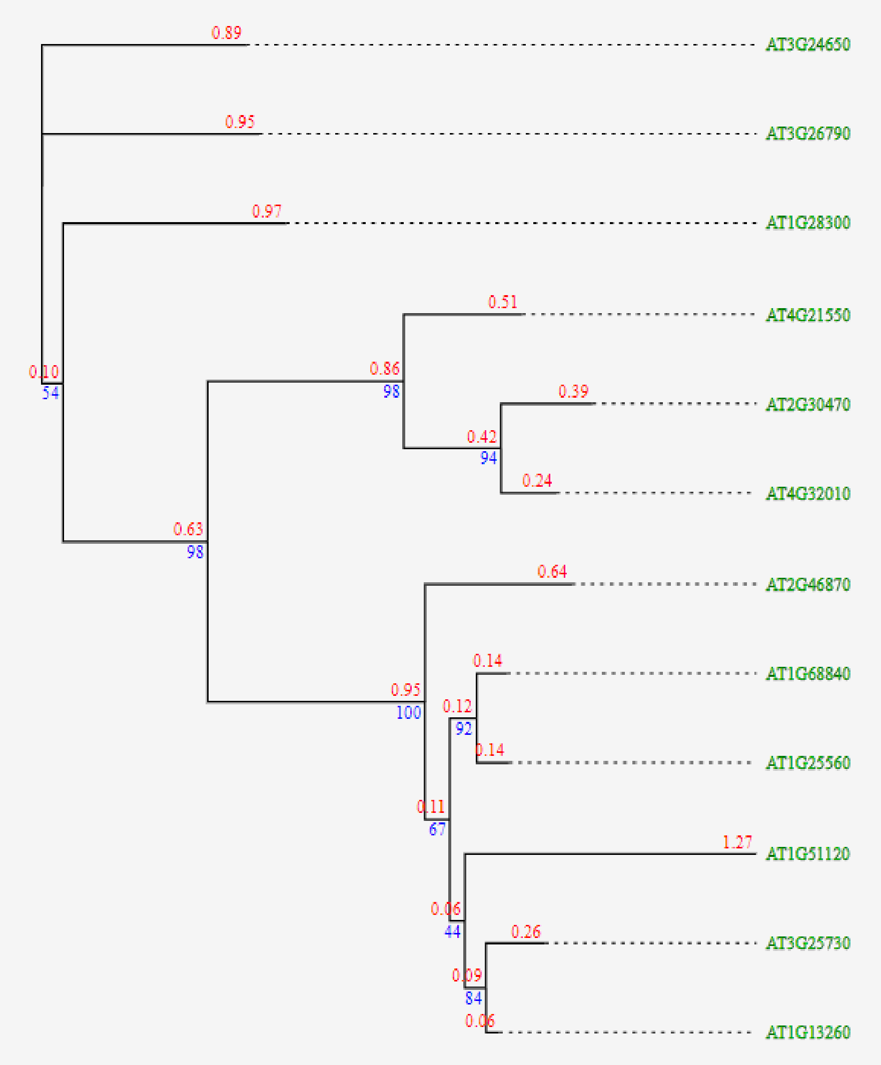


Figure 2: Custom PLAZA tree showing *ABI3* paralogs. Note that most of these aren't well-supported (bootstrap value of 54 for first the node on the left).

3.2.3 PANTHER

The PANTHER platform is possibly the most extensive online resource for the study of genes and proteins on a genome-wide scale. Initially aimed at classifying gene function using GO terms, it now provides a plethora of tools and data for studying gene function, evolutionary relationships, biochemical pathways, and/or analysis of gene expression or proteomics data (Mi et al., 2013). PANTHER has the additional advantage of being an active project that is updated monthly. Recently, a revised and detailed protocol on how to use PANTHER was published by its developers and we highly recommend consulting it (Mi et al., 2019).

3.3 Epigenomic Tools

3.3.1 EPIC-CoGe Browser for Arabidopsis epigenomic data

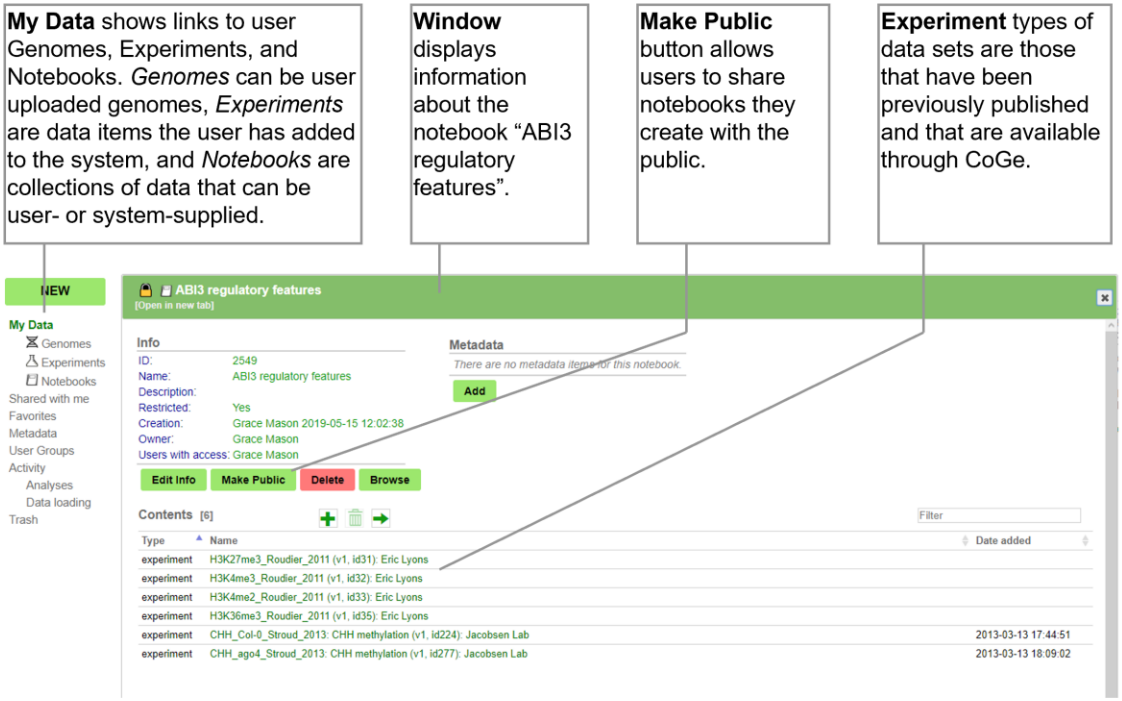
The EPIC-CoGe Browser (Nelson et al., 2018), built on JBrowse, is CoGe's visualization system for various types of genomic data. CoGe itself is short for "comparative genomics" and tens of thousands of genomes have been loaded into the system for easy comparison. EPIC is an acronym for the "Epigenomes of Plants International Consortium", a U.S. National Science Foundation-funded research coordination network (The EPIC Planning Committee, 2012). The power behind EPIC-CoGe is that it also functions as a massive data repository for hundreds of Arabidopsis sequencing experiments. If you choose to, you can upload your own data for visualization, add new genomes, share data easily with collaborators, as well as integrate your own data with publicly available experiments. To aid utility, searching for data sets of interest only requires knowing a keyword, such as "H3K4me3" or "Arabidopsis". You can also overlay several genomic data types by creating a notebook, as we will do in the steps below. For additional tutorials using EPIC-CoGe, please refer to https://genomevolution.org/wiki/index.php/EPIC-CoGe_Tutorial .

1. Log in to https://genomevolution.org/CoGe/User.pl using a free CyVerse account (instructions are given on the first page). After logging in, click on "My Data" on the left hand of page. Next, click on "Notebooks". At the top of this page, click on the plus sign, "+", to create a new notebook and name the notebook something informative. For this tutorial, we have named the notebook "ABI3 regulatory features". After designating a notebook name, click "Create Notebook". If you wish, you can also make the notebook publicly available by doing the following: double click on the notebook name to see a page describing the notebook and its attributes. Click on "Make public." Because notebooks can be shared publicly, all notebook names should be unique across all CoGe notebooks. That said, notebook names can be reused, however they will be assigned a unique ID number.

2. At the top of the page, use the search bar to find "CHH_ago4_Stroud_2013" (Stroud et al., 2013). Click the search result. On the right hand of the page under experiment details, click "add to Notebook". You will be prompted to search for your notebook name. After searching for the notebook to add this experimental data to, click "Add items".

3. Repeat these steps for the following data sets: “CHH_Col-0_Stroud_2013 (use id224), H3K27me1_Roudier_2011 (use id36), H3K27me3_Roudier_2011 (use id31), H3K36me3_Roudier_2011 (use id35), H3K4me2_Roudier_2011 (use id33), H3K4me3_Roudier_2011 (use id32) (Roudier et al., 2011; Stroud et al., 2013) .

4. In Step 3, we added several data sets together to create a combined genome annotation track that will display CHH-methylation data and ChIP-seq data for five different histone marks. **Figure 3** shows the attributes of the notebook and all of the data sets that it contains.



My Data shows links to user Genomes, Experiments, and Notebooks. *Genomes* can be user uploaded genomes, *Experiments* are data items the user has added to the system, and *Notebooks* are collections of data that can be user- or system-supplied.

Window displays information about the notebook “ABI3 regulatory features”.

Make Public button allows users to share notebooks they create with the public.

Experiment types of data sets are those that have been previously published and that are available through CoGe.

ABI3 regulatory features
[Open in new tab]

Info

ID:	2549
Name:	ABI3 regulatory features
Description:	
Restricted:	Yes
Creation:	Grace Mason 2019-05-15 12:02:38
Owner:	Grace Mason
Users with access:	Grace Mason

Metadata
There are no metadata items for this notebook.

Contents [6]

Type	Name	Date added
experiment	H3K27me3_Roudier_2011 (v1, id31): Eric Lyons	
experiment	H3K4me3_Roudier_2011 (v1, id32): Eric Lyons	
experiment	H3K4me2_Roudier_2011 (v1, id33): Eric Lyons	
experiment	H3K36me3_Roudier_2011 (v1, id35): Eric Lyons	
experiment	CHH_Col-0_Stroud_2013: CHH methylation (v1, id224): Jacobsen Lab	2013-03-13 17:44:51
experiment	CHH_ago4_Stroud_2013: CHH methylation (v1, id277): Jacobsen Lab	2013-03-13 18:09:02

Figure 3: EPIC-CoGe notebook attributes

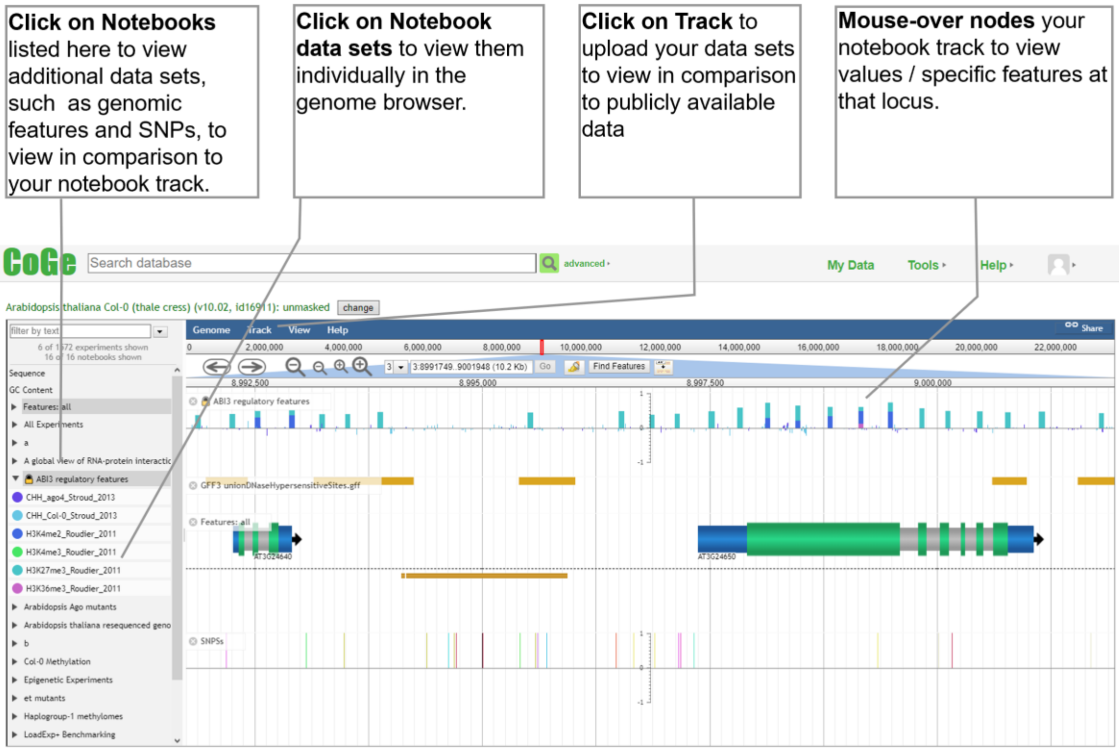
5. Next, you can view this combined track, or notebook, on the Arabidopsis genome browser supplied by CoGe and powered by JBrowse. Click “My Data” at the top right hand of the page. Next click Notebooks. You will see all the notebooks that you have created. Double click “ABI3_regulatory_features”, or your notebook’s name. Next, click “Browse.”

6. On the left hand of the page you should see a summary of your notebook along with notebooks created by other users that are available for browsing. In addition to your own notebook, you should also display “Features: all” and “SNPs” by clicking on these links. “Features: all” contains typical genomic features of *A. thaliana* including but not limited to: genes, transposons, small RNA loci, pseudogenes, *etc*. (Krishnakumar et al., 2015). “SNPs” displays data from the 1001 Genomes project that describes polymorphic loci across *A. thaliana* ecotypes (1001 Genomes Consortium, 2016).

7. Next, visualize the at the locus containing *ABI3*. At the top of the genome browser, enter the following coordinates: “3:8991671..9001200”. You should now be looking at *ABI3* and its upstream region.

8. You may also be interested in viewing your own data in this instance of JBrowse. To do this, click “Track” at the top of the browser and select “Open track file or URL”. From here select the file to upload, in this case we’ve provided an example data set: “unionDNaseHypersensitiveSites.gff3”[3] (Sullivan et al., 2015). Click “Open”.

9. In **Figure 4**, you can see the tracks that you combined in your notebook, the track that you uploaded, as well as the “SNPs” and “Features: all” notebooks. What you may notice is that there are quite a few SNPs in the upstream region of *ABI3* that also overlap a transposable element. Most interestingly, this region is also highly sensitive to DNaseI cleavage as evidenced by the union peak track indicating likely active transcriptional regulation. For the CHH-methylation data, we can see that the upstream transposons are depleted the for this methyl-cytosine context in the *ago4* mutant. There also appears to be a region that is highly enriched in H3K27me2 within the gene body of *ABI3*. You may even predict that some of the SNPs located in this putative regulatory region may have functional consequences for *ABI3* expression, a hypothesis which could be easily tested.



Click on Notebooks listed here to view additional data sets, such as genomic features and SNPs, to view in comparison to your notebook track.

Click on Notebook data sets to view them individually in the genome browser.

Click on Track to upload your data sets to view in comparison to publicly available data

Mouse-over nodes your notebook track to view values / specific features at that locus.

Figure 4. EPIC-CoGe Browser output for selected data sets, near the *ABI3* locus.

[3] http://bar.utoronto.ca/~nprovart/unionDNaseHypersensitiveSites.gff3

Two other useful sites that we call your attention to here are the Ecker Laboratory's 1001 Epigenomes Browser at http://neomorph.salk.edu/1001.aj.php (Kawakatsu et al., 2016) and the Jacobsen Lab Epigenomics Browser at https://www.mcdb.ucla.edu/Research/Jacobsen/LabWebSite/P_EpigenomicsData.php, where several publications' worth of data are available for browsing. However, EPIC-CoGe appears to be the most comprehensive resource to date.

3.4 Expression Analysis

Online expression analyses can be useful in place of performing Northern analyses, quantitative RT-PCR or constructing promoter:reporter fusions to determine patterns of expression. For instance, imagine we had identified an *abi3* mutation by positional cloning and wanted to know more about its biological function, and perhaps to guide us where to look elsewhere for a phenotype. One of the first steps would be to examine its expression pattern. Online tools such as the eFP Browser or Genevestigator makes this very easy, provided the platform used for measuring the transcriptome is able to detect the transcript for one's gene of interest[4].

3.4.1 eFP Browser

The eFP ("electronic fluorescent pictograph") Browser at the Bio-Analytic Resource for Plant Biology at http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi (Winter et al., 2007) provides easy access to 150 million expression measurements from *A. thaliana*, soybean (*Glycine max*), barrel medic (*Medicago truncatula*), poplar (*Populus trichocarpa*), maize (*Zea mays*), barley (*Hordeum vulgare*), rice (*Oryza sativa*) and others. Four-fifths of the measurements were made using *Arabidopsis* samples. Small pictographs are used to represent the experimental samples and contexts from which the expression data were generated, while differing expression levels within these samples are denoted by a colour scale.

1. Go to http://bar.utoronto.ca and select "*Arabidopsis* eFP Browser" from the BAR's homepage.

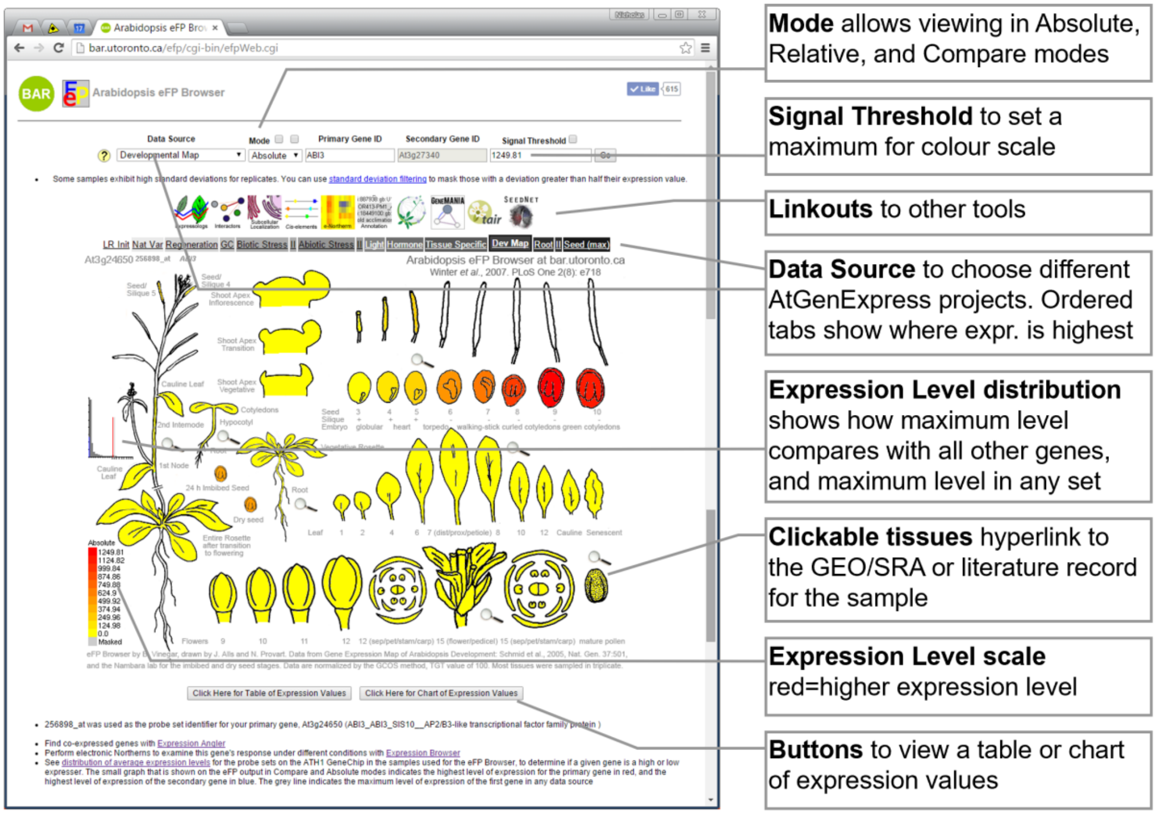
2. Enter your gene of interest's AGI ID[5] or name. In our case, we enter "At3g24650" or type "ABI3" for the *ABI3* gene into the Primary Gene ID box. Click Go.

3. **Figure 5** shows the output when querying the eFP Browser using *ABI3* with the default settings. The tissues that were sampled by Schmid *et al*. (2005) for their "gene expression map during *Arabidopsis* development" and by Nakabayshi *et al*. (2005) for the dry and imbibed seed samples are depicted in a pictographic manner. Where the expression (expression meaning steady-state mRNA levels) of *ABI3* is higher, the redder

[4] For instance, older microarray platforms are able to detect varying numbers of transcripts. Data from these are still quite useful and vast numbers of expression profiling experiments have been conducted with them. The ATH1 array from Affymetrix has probe sets for 22,814 transcripts, some of which may come from several genes. Next generation sequencing technologies, *i.e.* RNA-seq, are more comprehensive.

[5] The Arabidopsis Genome Initiative identifier, AGI ID, is easily found at TAIR.

is that tissue's colour. If there is little expression in a tissue, then it is coloured yellow.



Data Source: Developmental Map
Mode: Absolute
Primary Gene ID: ABI3
Secondary Gene ID: At3g27340
Signal Threshold: 1249.81

- Some samples exhibit high standard deviations for replicates. You can use standard deviation filtering to mask those with a deviation greater than half their expression value.

eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

- 256898_at was used as the probe set identifier for your primary gene, At3g24650 (ABI3_ABI3_SIS10__AP2/B3-like transcriptional factor family protein)

- Find co-expressed genes with Expression Angler
- Perform electronic Northerns to examine this gene's response under different conditions with Expression Browser
- See distribution of average expression levels for the probe sets on the ATH1 GeneChip in the samples used for the eFP Browser, to determine if a given gene is a high or low expresser. The small graph that is shown on the eFP output in Compare and Absolute modes indicates the highest level of expression for the primary gene in red, and the highest level of expression of the secondary gene in blue. The grey line indicates the maximum level of expression of the first gene in any data source

Mode allows viewing in Absolute, Relative, and Compare modes

Signal Threshold to set a maximum for colour scale

Linkouts to other tools

Data Source to choose different AtGenExpress projects. Ordered tabs show where expr. is highest

Expression Level distribution shows how maximum level compares with all other genes, and maximum level in any set

Clickable tissues hyperlink to the GEO/SRA or literature record for the sample

Expression Level scale red=higher expression level

Buttons to view a table or chart of expression values

Figure 5: eFP Browser output for *ABI3* in the Developmental Map view, showing strong expression in mature seeds.

4. By changing the Data Source, it is possible to explore other data sets that have been annotated in this pictographic manner. The eFP Browser also outputs where the expression of the gene of interest is strongest (in this case, in the Seed Data Source, not surprisingly, given *ABI3*'s known role there) but it is also worthwhile to examine other Data Sources[6]. For instance, *ABI3* also seems to be expressed in the vasculature tissue between the elongation and maturation zone of the root. If it had not already been known (Brady et al., 2003) that ABI3 is involved in root development, such an observation of expression in the root could guide us to look for phenotypes in the roots of *abi3* mutants more closely.

5. The Relative Mode option allows you to view expression of a given gene in each sample relative to its expression in a control sample, and to ascertain whether the gene's expression is above or below this level. If it is above, a red colour is used, and if it is

[6] It is useful to set the Signal Threshold to some value when comparing different genes or viewing a number of different Data Sources. That way, the expression level that "red" denotes is constant. The expression level distribution graph is also a handy feature for determining if one's gene of interest is has a strong level of expression. The small graph shows the distribution of the average expression level of all genes in the tissues depicted on the output, while the red line shows where the maximum expression level of the gene of interest falls along that distribution.

below a blue colour is used to colour the tissue in question. For the Developmental Map, this level has been computed as the median level across all of the tissues displayed. The Relative Mode is more useful in the case of “challenge” experiments, where a hormone or chemical has been applied as part of the experimental design. The control sample in this case would be the mock treated or untreated control.

6. If a given gene does not map to an ATH1 probe set, then try using the “Klepikova Atlas”, “Shoot Apex”, “Embryo”, “Silique”, or “Germination” Data Sources, as these were generated using RNA-seq.

3.4.2 eFP-Seq Browser

RNA-seq analysis can be thought of an extension of long-standing methods such as ESTs, SAGE, and MPSS (expressed sequence tags, serial analysis of gene expression, and massively parallel signature sequencing, respectively) for gene expression analysis. The main difference is that the overall number of “tags” that are generated for a given transcript population is far higher due to the efficiency of next generation sequencing machines at generating sequences cheaply, thereby increasing accuracy and sensitivity. The eFP-Seq Browser[7], included in the Bio-Analytic Resource for Plant Biology, is used to visualize this type of data. This tool allows us to search among 113 RNA-seq data sets used by Araport 11 to reannotate the Arabidopsis genome, and a collection of data sets from different organs and developmental stages published in Klepikova et al. (2016). The eFP-Seq Browser retrieves the number of reads mapped and display these above the desired Araport 11 gene model.

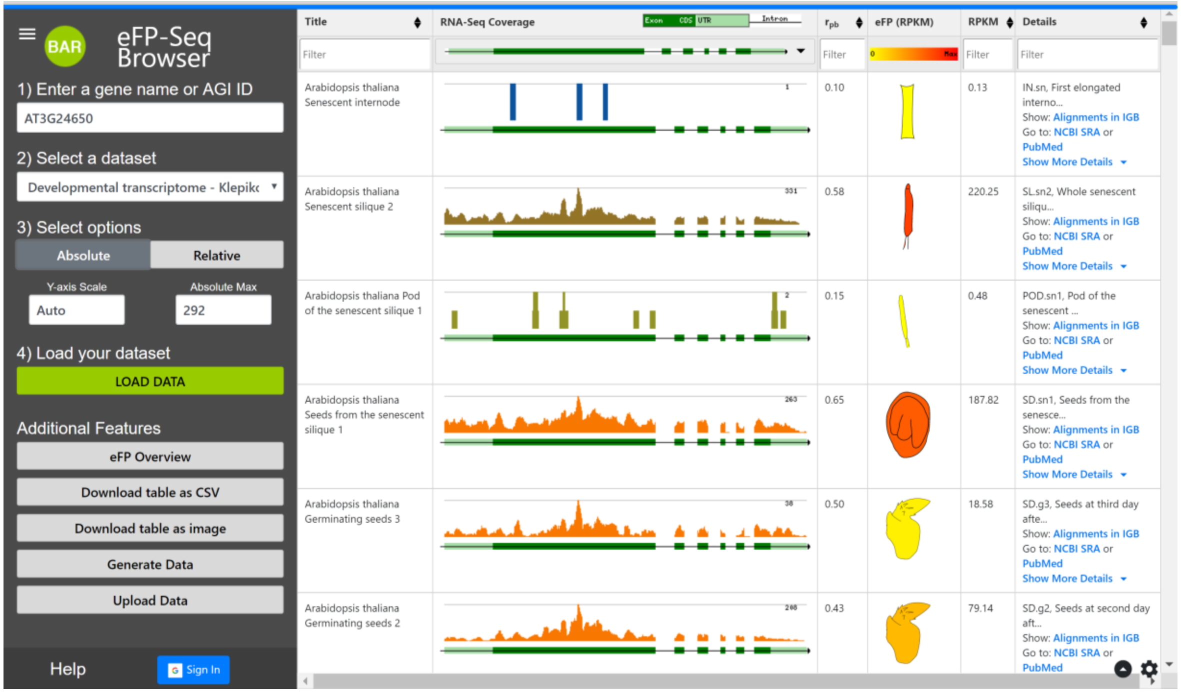
1. Go to http://bar.utoronto.ca and select “eFP-Seq Browser” from the BAR’s homepage.

2. Enter your gene of interest’s AGI ID. In our case, this is “At3g24650” for the *ABI3* gene. Click on the “Load Data” button.

3. **Figure 6** depicts the default output of this search. The data is presented in the same pictographic manner as the eFP browser. Additionally, the RNA-seq coverage is presented above the selected gene model variant. How well a given coverage profile maps to a gene model variant can help with the discovery of alternative splicing events.

4. Clicking on the RPKM column allows you to also sort your samples based on expression. Finally, data sets can be sorted or filtered by specifying cut-off values or key words on the headers of the columns. To identify readmap profiles that best match a given gene model, we can sort on the r_{pb} column from best score to worst.

[7] The eFP-Seq Browser paper by O’Sullivan et al. has been accepted at the Plant Journal



1) Enter a gene name or AGI ID: AT3G24650

2) Select a dataset: Developmental transcriptome - Klepikc

3) Select options

Y-axis Scale: Auto

Absolute Max: 292

4) Load your dataset

Title	RNA-Seq Coverage	r_{pb}	eFP (RPKM)	RPKM	Details
Filter		Filter		Filter	Filter
Arabidopsis thaliana Senescent internode	1	0.10		0.13	IN.sn, First elongated interno... Show: Alignments in IGB Go to: NCBI SRA or PubMed Show More Details
Arabidopsis thaliana Senescent silique 2	331	0.58		220.25	SL.sn2, Whole senescent siliqu... Show: Alignments in IGB Go to: NCBI SRA or PubMed Show More Details
Arabidopsis thaliana Pod of the senescent silique 1	2	0.15		0.48	POD.sn1, Pod of the senescent ... Show: Alignments in IGB Go to: NCBI SRA or PubMed Show More Details
Arabidopsis thaliana Seeds from the senescent silique 1	263	0.65		187.82	SD.sn1, Seeds from the senesce... Show: Alignments in IGB Go to: NCBI SRA or PubMed Show More Details
Arabidopsis thaliana Germinating seeds 3	38	0.50		18.58	SD.g3, Seeds at third day afte... Show: Alignments in IGB Go to: NCBI SRA or PubMed Show More Details
Arabidopsis thaliana Germinating seeds 2	268	0.43		79.14	SD.g2, Seeds at second day aft... Show: Alignments in IGB Go to: NCBI SRA or PubMed

Figure 6: eFP-Seq view of expression pattern of *ABI3* (*At3g24650*) in Arabidopsis. Stronger expression is denoted by larger RPKM values and darker colouration. Much like the eFP Browser, the interface provides many options for exploring the expression data, including filters. Density of reads across different exons can help with identifying different alternative splicing events in different samples.

3.4.3 TraVA

Although limited in content, the TraVA webpage presents an intuitive and straightforward interface to study the expression of our gene of interest. It contains a comprehensive spatiotemporal data set for *Arabidopsis thaliana*, encompassing RNA-seq data across 79 organs and developmental stages (Klepikova et al., 2016).

1. Go to http://travadb.org/browse. Make sure Arabidopsis is selected in the dropdown menu at the top of the screen.

2. Type in the AGI ID of your gene of interest next into the input box, at the top left of the page, *e.g.* At3g24650 for *ABI3*.

3. By default, the values shown in the boxes represent read counts normalized by method applied in median-of-ratio method as described in Anders and Huber (DESeq/DESeq2) and divided by maximum value of expression level, so all values vary from 0 to 1. As expected, in the case of *ABI3*, the largest relative values correspond to different stages of seed development. We can adjust and select different types of normalization and whether we prefer count value by clicking the small boxes on the right side.

4. We can then select a specific sample by clicking on top of the box containing its count values. Once selected, TraVA automatically changes the values shown to fold changes in regard to the selected sample. This allows us to easily visualize which samples are significantly different to our reference sample. By default, the statistically significant differences are calculated using DESeq. However, we can choose other algorithms such as DESeq2 or BaySeq by selecting them using the boxes on the right side.

3.4.4 Genevestigator

Data from 10,000+ high-quality ATH1 arrays are available for *Arabidopsis* from Genevestigator (https://www.genevestigator.com/gv/), see Hruz et al. (2008). As with the eFP Browser, the different tools of this resource let us determine when and where our gene of interest is expressed and in response to which conditions. The main difference between the eFP Browser and Genevestigator is that data are displayed in heatmap format as opposed to a pictograph. One of the major advantages of this tool is the simultaneous analysis of hundreds or thousands of genes in a biological context, as opposed to the eFP Browser, which permits a user to examine only one gene at a time[8].

1. Go to https://www.genevestigator.com/gv/ and select “Plant Biology” under “Application areas”. Click on “INSTALL APP” on the subsequent page. Follow the instructions to get Genevestigator running on your computer – will need to have the correct Java version, and also a username/password (plus, request free access to Genevestigator *Professional* at the bottom of your user account page).

2. Click “New” in the Data Selection panel on the left then choose “*Arabidopsis thaliana”* as the Organism, “Affymetrix Arabidopsis ATH1 Genome Array” as array type, and don’t select any Filters[9]. Click OK.

3. Introduce the AGI ID by clicking on “New” in Gene Selection panel of the main window. In our case, we enter the *ABI3* AGI ID, “*At3g24650*”. Click OK.

4. The Condition Search tools gives us gene expression data from the different array sets[10], the filled dots indicate detection *p*-values under 0.06 and the unfilled *p*-values over 0.06[11]. On the “Samples” tab we can examine the expression in all the available arrays. To get the experimental design and gene expression information, just move the mouse over the sample name or the dot.

[8] The Bio-Analytic Resource does provide a bulk query tool called “Expression Browser” which provides a Genevestigator-like ability to query many genes at one, see http://bar.utoronto.ca/affydb/cgi-bin/affy_db_exprss_browser_in.cgi.

[9] Genevestigator has no control over experimental design, only a post-hoc analysis is possible to check the quality of the array. For more information about quality control criteria visit https://genevestigator.com/gv/file/GENEVESTIGATOR_UserManual.pdf.

[10] For experimental normalization, Genevestigator uses Bioconductor's RMA implementation.

[11] A *p*-value under 0.06 indicates that the signal is reliably detected.

5. Click the Home icon to return to the overview screen. Click on the different tabs to explore the ontologies of anatomy and perturbations (including response in mutants; you will need to sign up for a free trial for this). The expression of *ABI3* is high in the seed arrays, principally in the embryo and endosperm, rather than in the seed coat. By genotype, *ABI3* is highly expressed in the pER8:LEC1 over-expression line and repressed in *lec1.1* plants; in contrast *ABI3* has lower expression in the *pif1/pif3/pif4/pif5* quadruple mutant plants. ABA treatments promote its expression, as does the treatment with Paclobutrazol (a GA inhibitor).

6. We can generate hypotheses from these data: phytochrome-mediated light signaling and downstream factors regulate *ABI3* expression, and LEC1 likely regulates *ABI3* expression either directly or indirectly**.**

7. Note: Genevisible at https://genevisible.com/search may be freely used to search the developmental or perturbation compendia Genevestigator has compiled to identify the 20 developmental data sets or conditions where your gene of interest has the strongest and weakest expression.

3.5 Coexpression Tools

Coexpression analysis can leverage the large number of gene expression data sets that have been generated in the past decade to answer the question “which genes show similar patterns of expression as my gene of interest, across all samples in a given database?”. Those that show similar patterns of expression may be involved in the same biological process as the query gene, after the “guilt-by-association” paradigm. The use of such analyses is well-covered in a review by Usadel and colleagues (2009).

3.5.1 Expression Angler

Expression Angler (Toufighi et al., 2005) is a powerful yet easy-to-use tool for identifying coexpressed genes, as measured by the Pearson correlation coefficient – *r*, in both a condition-dependent and condition-independent manner[12]. With it, it is possible to answer the question of which genes show similar patterns of expression in 9 different compendia – genes with an *r*-value of greater than around 0.75 can be considered coexpressed. It is also possible to use just a subset of the samples within a given data set to perform the analysis, which we will do below for *ABI3*. Those genes annotated as “unknown function” or those with vague descriptions may be involved in the same process as the query gene.

[12] It is often useful to examine condition-dependent data sets, as genes may respond one way in a set of tissues and in an opposite way in others. If one lumps these sets together, then these correlations cannot be detected. This issue is described in greater detail in the Usadel et al. (2009) review.

1. Go to the Bio-Analytic Resource for Plant Biology's homepage at http://bar.utoronto.ca and select the Legacy Expression Angler link (a revamped version of this tool is also available but it is difficult to do the following with it).

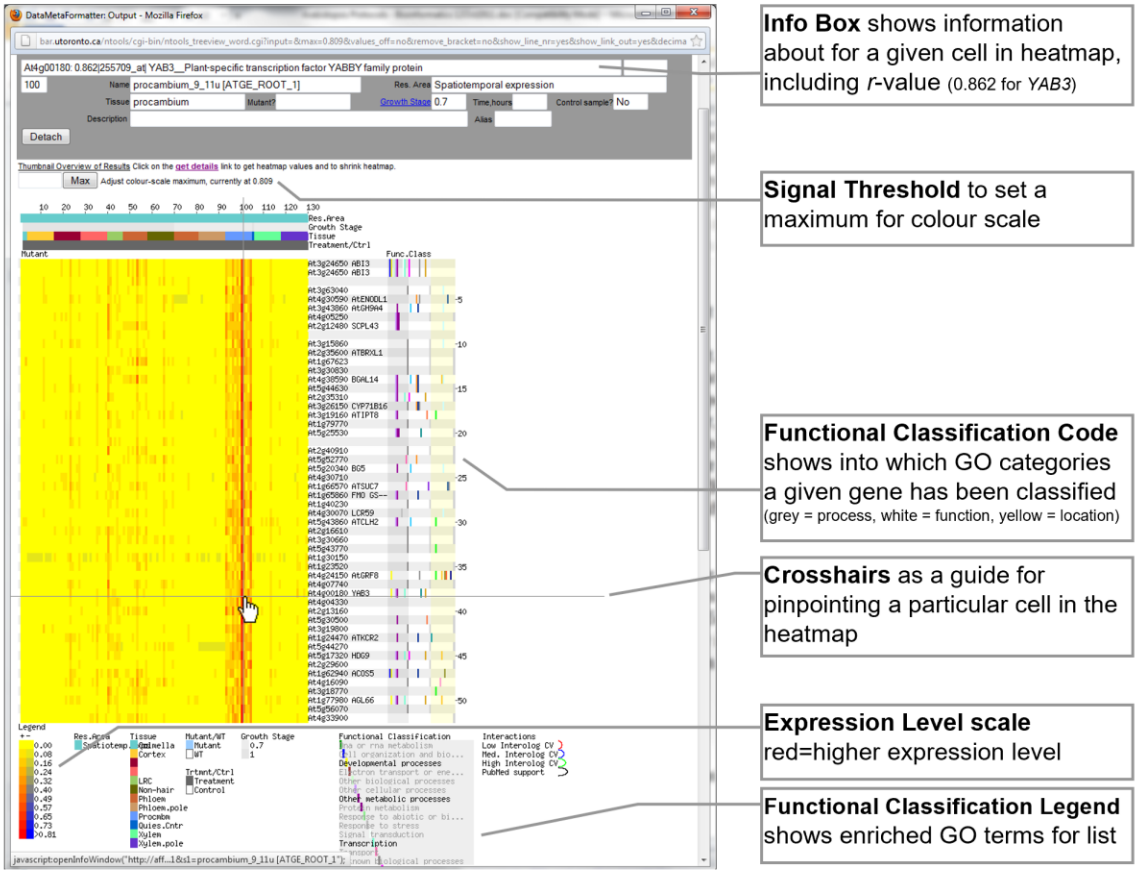
2. In normal use, select a data set and enter the AGI ID of interest. If we had used the AtGenExpress Tissue Set, which corresponds to the data set shown in **Figure 5**, we would identify many other seed maturation genes and ABA-responsive genes being coexpressed with *ABI3* – the top 50 of these are listed in **Supplemental Table 1**. Another way to use Expression Angler, however, is to define a subset of samples in which to search. Use the "Subselect and Custom Bait Page" link, and then choose a data set. In this case we will use the "Root Compendium". On the input page, we will enter *At3g24650* and then select "Return just the top 50 hits" in the 129 samples of the "Spatiotemporal expression" experiment (Brady et al., 2007)[13].

3. Click "Submit Query" at the bottom of the page.

4. On the output page, examine the "View formatted data set after median centering and normalization", as shown in **Figure 7**. This view is closest to the way that Expression Angler finds expression pattern similarity with the Pearson correlation coefficient, which standardizes gene expression values by the average value (not median) when comparing two expression vectors. Another useful display is the "View formatted data set", which shows untransformed expression levels.

5. By mousing over the heatmap, it is possible to find out the annotation of the genes, which samples they are expressed most strongly in, and other information. Interestingly, *YABBY3*, likely a patterning gene, shows up as being coexpressed with *ABI3*, as are several other transcription factors.

[13] Given the number of samples in most of these data sets, even a Pearson correlation coefficient of 0.3 can be considered "significant". But with this *r*-value, only $(0.3)^2$= 9% of the variance is shared between two genes. An *r*-value of 0.7 means that coexpression explains 49% of the variance in common between two genes. This is the reason why 0.7 to 0.75 is often used as a cut-off for coexpression analysis.



DataMetaFormatter: Output - Mozilla Firefox

bar.utoronto.ca/ntools/cgi-bin/ntools_treeview_word.cgi?input=&max=0.809&values_off=no&remove_bracket=no&show_line_nr=yes&show_link_out=yes&decima

At4g00180: 0.862|255709_at| YAB3__Plant-specific transcription factor YABBY family protein

100 Name: procambium_9_11u [ATGE_ROOT_1] Res. Area: Spatiotemporal expression

Tissue: procambium Mutant?: Growth Stage: 0.7 Time,hours: Control sample?: No

Description: Alias:

Detach

Thumbnail Overview of Results Click on the get details link to get heatmap values and to shrink heatmap.

Max Adjust colour-scale maximum, currently at 0.809

Info Box shows information about for a given cell in heatmap, including *r*-value (0.862 for *YAB3*)

Signal Threshold to set a maximum for colour scale

Functional Classification Code shows into which GO categories a given gene has been classified (grey = process, white = function, yellow = location)

Crosshairs as a guide for pinpointing a particular cell in the heatmap

Expression Level scale red=higher expression level

Functional Classification Legend shows enriched GO terms for list

Figure 7: Heatmap output of Expression Angler after searching in the Root Spatiotemporal data set of Brady *et al*. (2007) with *ABI3*.

3.5.2 ATTED II

ATTED II (Aoki et al., 2016) is a gene coexpression database for finding functional relationships between genes. This tool uses the mutual rank (MR) of the Pearson's correlation coefficient (Obayashi and Kinoshita, 2009) to investigate gene coexpression in *Arabidopsis* in a condition-independent way or across five sets of experimental conditions: tissue, abiotic stress, biotic stress, hormones and light conditions. ATTED II also offers analysis of rice coexpression data to provide a comparative view between both species using putative gene orthologs.

1. Go to http://atted.jp/.

2. On the search menu, click on the arrow(s) on the right-hand side of the pull-down menu and select the option that best fits your search ("All words", "Keyword", "Gene alias", "Gene ID" or "GO ID"). We will search by "Gene ID", *At3g24650* for *ABI3.* Click Search after entering it in the box.

3. The output window shows a brief description of the gene of interest, such as the alias and the function. By clicking the little "L"-shaped icon in the Locus Page column, ATTED II sends us to a new window with a lot of information about the gene: functional annotation, a gene coexpression network, gene expression levels and predicted *cis* elements.

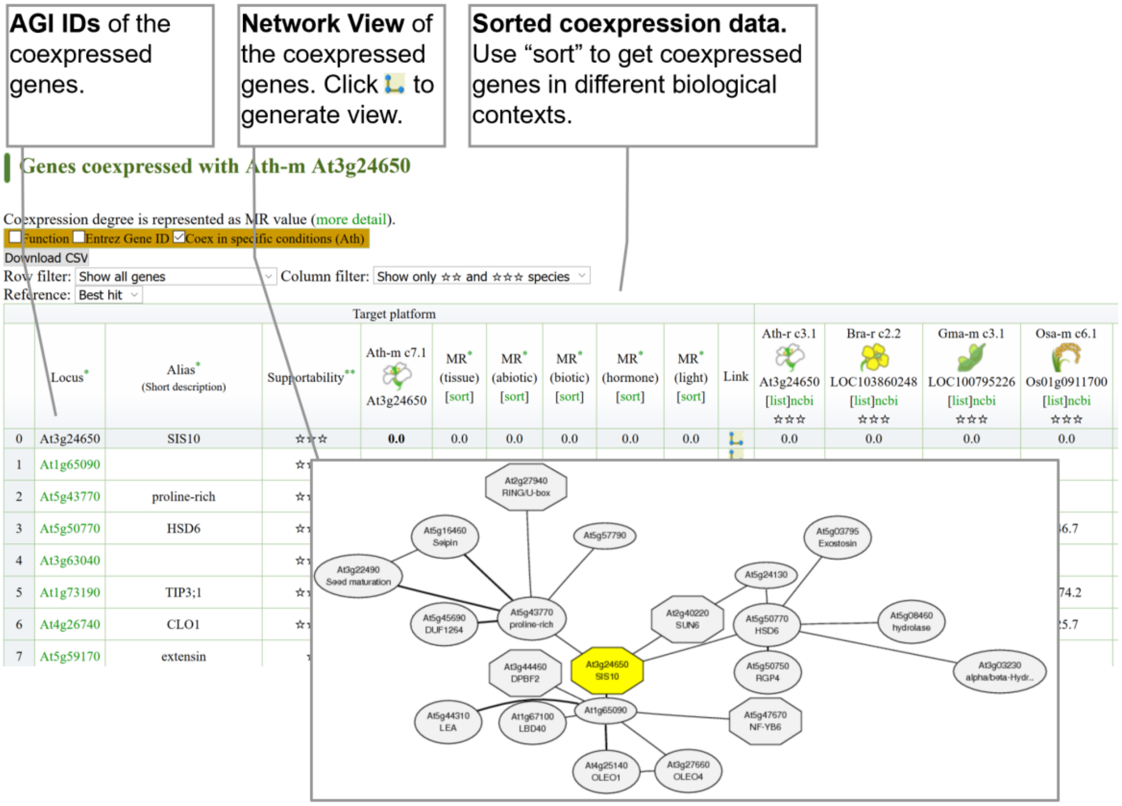
4. For a more extensive analysis of coexpressed genes, go back to the locus search window and click on "list" of coexpressed genes. The program will give a list of the top 300 coexpressed genes[14].

5. Check "coex in specific conditions" to study coexpression under different conditions: tissue, abiotic stress, biotic stress, hormone and light. We can rank coexpression in each condition by clicking on "sort" in that column's heading. This approach would help us to infer the gene's function in each category. For instance, the genes that are more closely correlated to *ABI3* differ extensively depending on which biological context in which we are interested. This suggests that *ABI3* has multiple functions – both developmentally and in response to the environment, *i.e.* if we sort by "tissue", *ABI3* is coexpressed with several seed-associated genes, whereby different genes show up at the top of the *ABI3*-coexpressed lists under hormone treatments or abiotic stress.

6. Click on the small "L"-shaped icon in the Link column for each coexpressed gene to get the same information described in step 3. One of the most powerful features of ATTED II is the network visualization of coexpressed genes. This network visually depicts genes connected directly and indirectly to our query gene by coexpression. We can explore coexpression network neighbourhoods by clicking on the gene-names, see **Figure 8**.

7. ATTED II shows that *ABI3* is coexpressed with *EPR1* (an extensin-like gene) that is involved in seed germination, but only expressed in the endosperm (Dubreucq et al., 2000). *AIL5* (*AINTEGUMENTA LIKE-5*; also called *PTL5*) appears to be coexpressed with *ABI3* as well. *AIL5* encodes a member of the AP2 family of transcriptional regulators that are involved in cell proliferation activities in many organs (Nole-Wilson et al., 2005). *AIL5* mutants are resistant to ABA. We can therefore hypothesize that ABI3 and AIL5 act together to control cell proliferation and/or ABA response.

[14] ATTED II uses the MR (Mutual Rank) value to rank the co-expressed genes, lower MR values means more correlation. This method was determined by the authors to have higher performance in the prediction of gene function than the Pearson Correlation coefficient (PCC).



AGI IDs of the coexpressed genes.

Network View of the coexpressed genes. Click to generate view.

Sorted coexpression data. Use "sort" to get coexpressed genes in different biological contexts.

Genes coexpressed with Ath-m At3g24650

Coexpression degree is represented as MR value (more detail).

□ Function □ Entrez Gene ID ☑ Coex in specific conditions (Ath)

Download CSV

Row filter: Show all genes

Column filter: Show only ☆☆ and ☆☆☆ species

Reference: Best hit

				Target platform										
	Locus*	Alias* (Short description)	Supportability**	Ath-m c7.1 At3g24650	MR* (tissue) [sort]	MR* (abiotic) [sort]	MR* (biotic) [sort]	MR* (hormone) [sort]	MR* (light) [sort]	Link	Ath-r c3.1 At3g24650 [list]ncbi ☆☆☆	Bra-r c2.2 LOC103860248 [list]ncbi ☆☆☆	Gma-m c3.1 LOC100795226 [list]ncbi ☆☆☆	Osa-m c6.1 Os01g0911700 [list]ncbi ☆☆☆
0	At3g24650	SIS10	☆☆☆	**0.0**	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0
1	At1g65090		☆☆											
2	At5g43770	proline-rich	☆☆											
3	At5g50770	HSD6	☆☆											6.7
4	At3g63040		☆☆											
5	At1g73190	TIP3;1	☆☆											74.2
6	At4g26740	CLO1	☆☆											5.7
7	At5g59170	extensin	☆											

Figure 8: Output of an ATTED II query for the *ABI3* gene, showing ranked list of coexpressed genes in ATTED II's condition independent data set (top panel) and a visualization of the coexpression list in network form (insert).

3.5.3 AraNet

AraNet uses machine learning to train networks of correlated genes from a "gold standard" of reliable Gene Ontology annotations and pathways from MetaCyc, using the guilt-by-association principle and log-likelihood scores. While it uses data other than expression data to find associations, a lot of the data it does use are in fact expression data sets and as such we are including AraNet in the coexpression tool category.

1. Go to https://www.inetbio.org/aranet. Click on "Network-search". On the following page, click "Query Option 1. Find new members of a pathway".

2. Under "Gene Set", enter "AT3G24650". You can leave Organism on "Auto-Detection". Click "Submit".

3. You will need to enable Flash on the next page to see the interactive views (that are powered by Cytoscape Web – in Chrome click on the "Secure" beside the URL to "Allow"). The first view becomes more interesting with multiple genes. Scroll down to the

one element table entry on “AT3G24650”. You will note that the associated GO terms for *ABI3* seem to make sense.

4. Keep scrolling to the second interactive view (try viewing in a new window if you can’t see anything). You can see how many genes AraNet associates with *ABI3* (SIS10) by counting how many entries are in the table.

5. A nice feature of AraNet is use it to look for associations in common between genes of interest. Start a new “Network-search” → “Find new members of a pathway”. Under “Gene Set”, this time enter both *ABI3* “At3g24650” and *LEC1* “At1g21970”. You can leave Organism on “Auto-Detection”. Click “Submit”.

6. Once again, scroll down to the second network view. Take a look at the node labeled “FG”. You will see that it is associated with both *ABI3* and *LEC1*. In this way, we can use AraNet to help us understand connections between genes or gene products.

7. We can also use the “Query option II. Infer functions from network neighbors” under “Network-search” to try to infer functions based on the GO annotations of the genes associated with our gene or genes of interest – check it out!

3.5.4 AtCAST2

Instead of looking for genes with similar expression patterns across a set of samples, AtCAST allows you to ask questions about correlations between experiments, that is, it performs the correlation calculation by generating a vector across all genes for a given experiment and then asking “is there another sample where those genes also show similar signatures of expression”. It does this for all data sets in its database, but importantly, allows you to also ask for your own data set, perhaps generated from a mutant-of-unknown-function line or a chemical-of-unknown-mode-of-action treatment, “what data set in the public databases most closely resembles mine?” It is kind of like Blast for a transcriptomics experiment you’ve done, except instead of asking what sequences are similar, you are asking what experiments are similar, based on their transcriptomic profiles!

1. Go to http://atpbsmd.yokohama-cu.ac.jp/cgi/atcast/home.cgi. Click on AtCAST 3.1. Click on the “> Analyze Public Data” button about halfway down the page. Under “Find experimental condition by keyword”, enter “seed” and click “Search”.

2. Click on the result labelled “Seeds stage8 w/o siliques”. Using the eFP Browser in Section 3.4.1, we saw that one of the highest level of expressions for *ABI3* was seen here. Scroll down and take a look at the data visualization, seen in **Figure 9**.

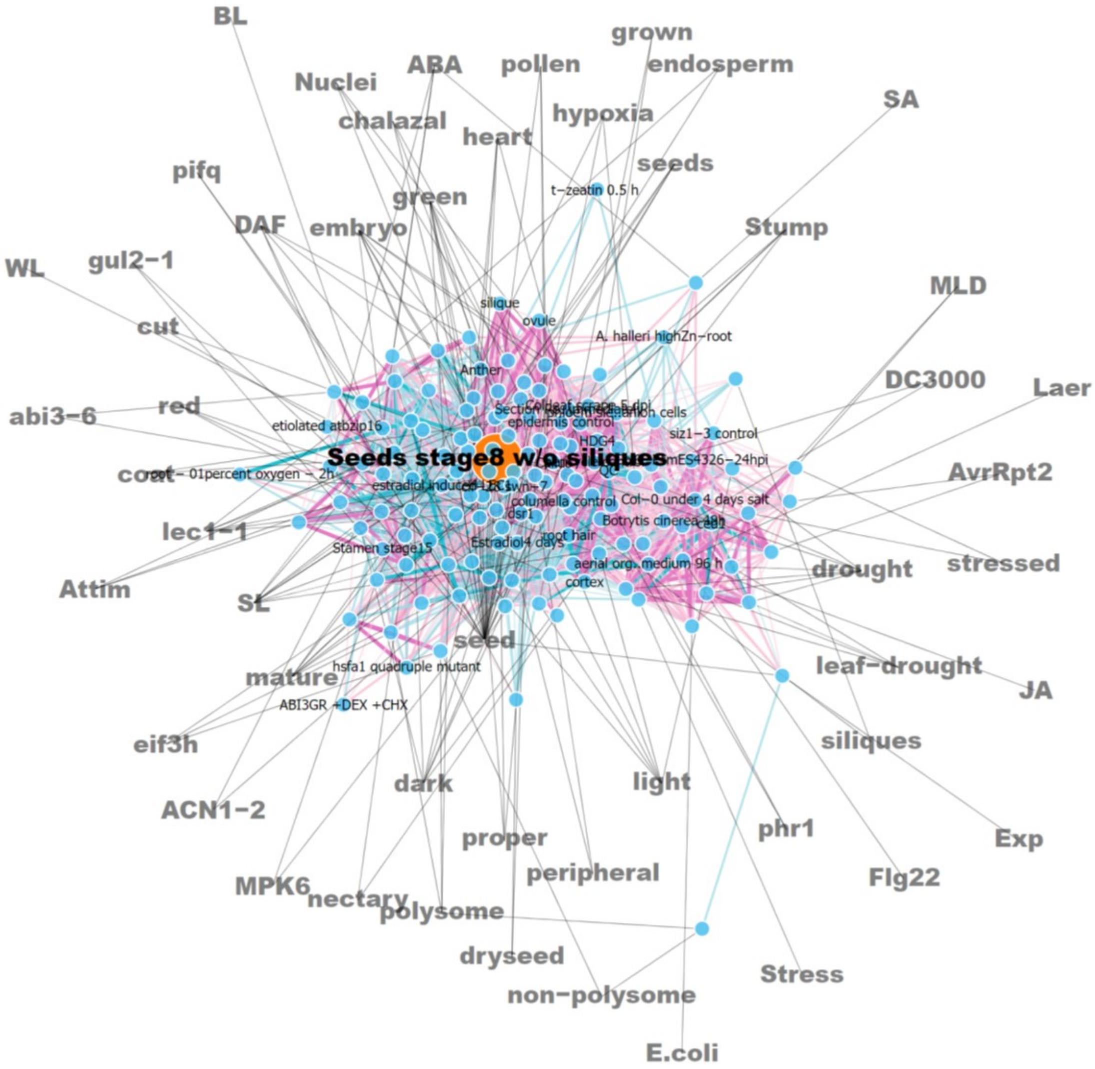


Figure 9: AtCAST visualization of samples correlated with Stage 8 seeds at the transcriptomic level. Pink edges denote positive correlation.

3. Continue scrolling until you get to the table labeled “List of experiments…” This table compares the overall expression pattern of the selected experiment with the other experiments in the database. We are interested in *ABI3*, and there happens to be an *ABI3* knockout experiment: scroll down to “abi3-6 16 DAF Seed”. Notice how the correlation is negative. Click the leftmost “More info” link under the “Correlation from…” columns. You will see a scatter plot showing the expression of differentially expressed genes in the two experiments. It is pretty amazing to see how anti-correlated this data set is.

We hope you can appreciate how powerful this tool would be if you had a mutant whose mode-of-action was unknown. Simply by analyzing your own expression data from that mutant with AtCAST you would quickly be able to identify similar experiments in the database, which could lead you quite rapidly to a functional hypothesis as to the role for the mutant gene.

3.6 Promoter Analysis

Gene expression is dependent on the *cis*-regulatory elements present in the promoter regions of genes. These elements act as binding sites for one or more transcription factors. Many tools have been developed to better understand how these transcription factor-binding sites might regulate such expression. In this section we will introduce tools that will help us to analyze and visualize promoter regions of *Arabidopsis* genes.

3.6.1 Cistome

Imagine a set of genes that are coexpressed in response to a certain stimulus. It will be of interest to determine common upstream regulatory motifs between these genes that could explain this particular behaviour and identify putative upstream regulators. Cistome (Austin et al., 2016) is a tool that searches for enriched motifs in the promoter regions of these genes.

1. Go to http://www.bar.utoronto.ca/cistome/cgi-bin/BAR_Cistome.cgi. Enter the AGI ID list in the “Enter a list of genes” box and click “Add to List”. You will use the top 50 coexpressed genes for ABI3 across a “Developmental Map” as identified with the Expression Angler tool – **Supplemental Table 1**.

2. Choose “TSS/TrSS (TAIR upstream)” as the start position and 1000 bp as the sequence length.

3. Choose a motif set. In this section, we are interested in studying whether a particular motif is over-represented in the promoter regions of our gene set. Under the “Enter Motifs” tab, select “Paste in your own PSSMs or consensus sequences” (PSSMs are Position Specific Scoring Matrices, a more flexible way to represent transcription factor binding sites and describes the probability of how often a given nucleotide can be present at each position of the motif). Select the blank option for the “Data set” dropdown menu.

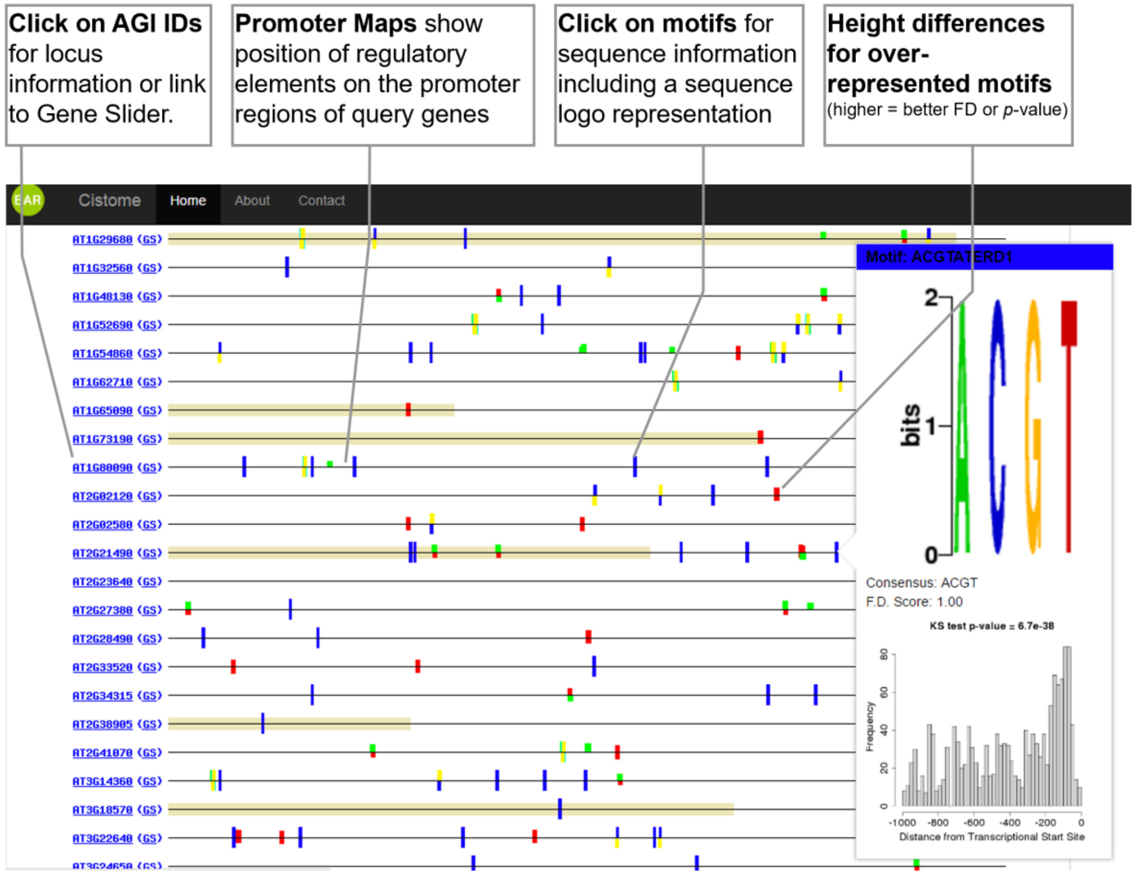
4. Enter the search sequence in the format required. Here, we will use the G-box motif (CACGTG), which is a binding site for the PIF transcription factor family.

5. Tick the Consensus Sequence option. We will enter the motif sequence in Fasta format, over two lines – copy the following to do so, including a return after the “CACGTG” motif
   ```
   > GBOX
   CACGTG
   ```

6. Toggle “Only show significantly enriched motifs (slow)”. You can also specify the significance parameters. In this example analysis, we will use the default parameters, which includes a Z-score cutoff of greater than 3, a functional depth cutoff of 0.35 and that this motif must be found in at least half of the genes in the gene set.

7. Perhaps you are also interested in searching additional motifs. To search for known motifs, return to the input page and keep all other settings the same but choose “Only *Arabidopsis* PLACE elements” in the Step 3 “Data set” dropdown menu, which will use one of two parts of a previously published motif database, PLACE (Higo et al., 1998), which contains around 100 cis-elements from plants, manually curated from published, small-scale studies (the G-box motif is encompassed in this set).

8. Click on “Begin Search” and Cistome will display a diagram with the over-represented regulatory elements mapped on the promoters of the genes included in the analysis (this analysis may take 2-3 minutes; be patient), see **Figure 10**. Cistome determines over-representation by comparing the frequency of occurrence of each motif against the frequency of occurrence of the same motif in randomly selected sets of promoters from the background set. We set that the G-box is indeed over-represented in our set of promoters, suggesting our hypothesis regarding PIF-family transcription factor regulation is correct.



Click on AGI IDs for locus information or link to Gene Slider.

Promoter Maps show position of regulatory elements on the promoter regions of query genes

Click on motifs for sequence information including a sequence logo representation

Height differences for over-represented motifs (higher = better FD or *p*-value)

Figure 10: Partial output of a Cistome query that represents the over-represented regulatory elements mapped onto the promoters of *ABI3* developmentally coexpressed genes. Blue symbols represent the ERD1 motif. The G-Box motif is also returned as being significantly over-represented in this set of genes.

9. Some other useful aspects of the Cistome tool: click on “Cluster View” at the tab along the top of the Cistome output. Cistome will displays a dendrogram of the

overrepresented motifs based on the similarity of the PSSMs generated from the mapping procedure.

10. Click on “Seq Logo View” to get the frequency of the distinct nucleotides that are found in the over-represented binding sites. Once you have a given sequence motif you can identify other genes in the genome that may contain this element. You can then query coexpression databases to see if these genes are coexpressed with your gene of interest or, in this example, if they are coexpressed with *ABI3* under any other conditions. This would suggest common regulatory control of a suite of functionally related genes.

3.6.2 MEME-suite

Similar to Cistome, tools within MEME-suite (Bailey et al., 2009) can search a collection of genes for enriched sequences, or scan sequences of interest to find matches to known motifs. Unlike Cistome, users can take advantage of the DAP-seq data set from the Ecker Lab, which has identified transcription factor binding sites for hundreds of transcription factors (O’Malley et al., 2016). Users can also perform *de novo* motif enrichment to discover new motifs independent of a motif database. Here, we will discuss how to utilize two tools from MEME-suite: FIMO (Grant et al., 2011) and AME (McLeay and Bailey, 2010). FIMO (Find Individual Motif Occurrences) scans one or more sequences for individual matches to each motif in a database (see Supplemental Data from O’Malley et al., 2016). AME (Analysis of Motif Enrichment) searches for known motifs that are enriched across input sequences compared to control sequences.

3.6.2.1 FIMO

1. Go to http://meme-suite.org/tools/fimo.

2. Upload the supplied file of DAP-seq motifs (ArabidopsisDAPv1.meme, available here[15]) under the “Input the motifs” menu. If you open the database file in a text editor, you will observe PSSMs for ~800 transcription factors assayed by DAP-seq. Again, PSSMs represent transcription factor binding sites and describe the probability of how often a given nucleotide can be present at each position of the transcription factor motif.

3. Under “Input primary sequences”, click the drop-down menu and select “Type in sequences”. Copy and paste the following (FASTA format):

```
>ABI3_upstream_region
atgtctttctctcgaggaactttgtttttatttcttagaagatgaggggagatttactatctaaa
taaaattttaaatgtttgtaagtattatgagctcaacaattttgtcaatagtgccacaaatttaa
acgtttgctttttgtcttctttgaaaaatcaaatgctgaaaaactgttacatctctttttcttaa
aaactcttgtctctctccttttccttctgctgaggtaattgaatgctgcaaagagaaagagaata
acttaaacccaaattacacttaccgccagaaaaaaaaaaagagttcagtttaatctaacatattt
tatacaatacaattgaattatattagtaaaaaaaaacttccatataaatcatggaacaaactgg
```

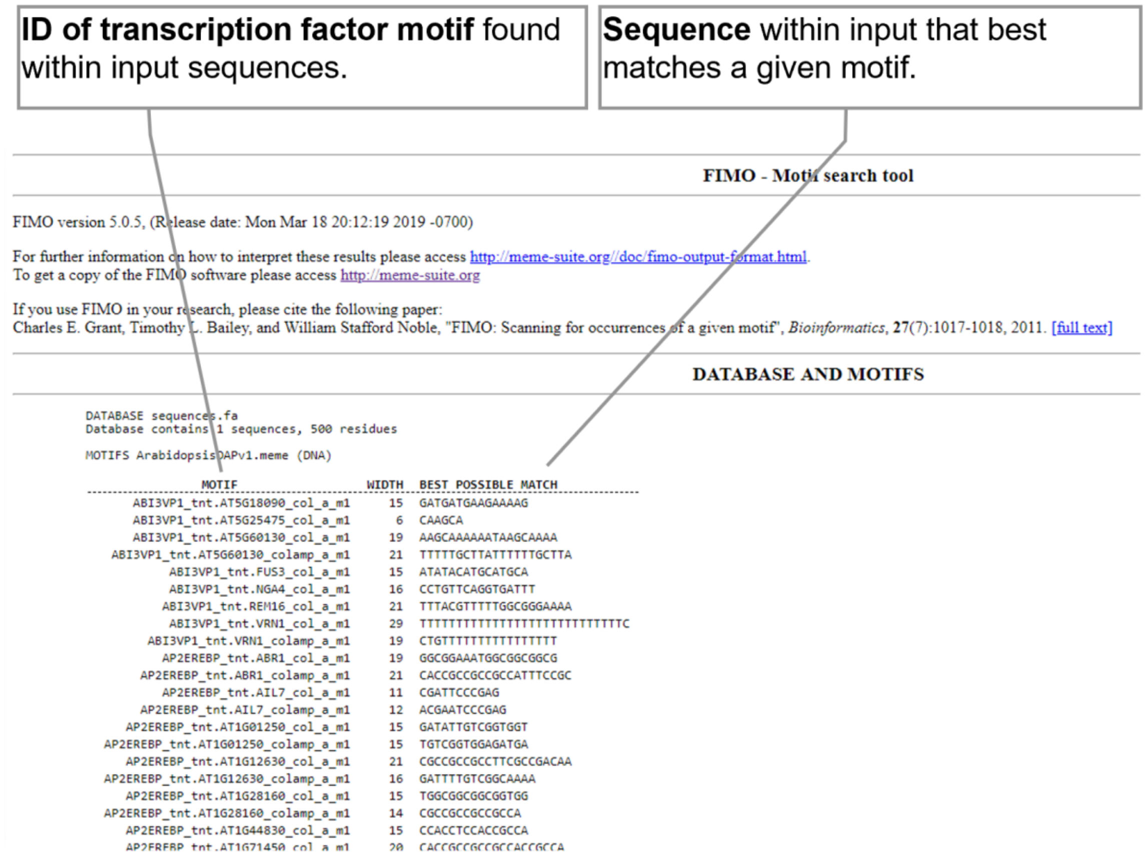
[15] Available at http://bar.utoronto.ca/~nprovart/ArabidopsisDAPv1.meme

```
aacacatgggctctcttatttaatttattttctttttgagggatttaaccatgtttattatat
agttttataaatatatatataccatctctccataatttataaaat
```

4. Under “Enter in job details”, type in your email address and as well as a job description to remind you of what you submitted. This step may take more than an hour to run on the MEME-suite server, so please be patient.

5. When the job is done, you should receive an email with a link to your results. Click the link and you will be taken to an output page that has the following links: FIMO HTML output, FIMO TSV output, Input Sequences, Calculated Alphabet, and Calculated Background. For future data analysis, download the TSV file. If you open this TSV file using a text editor, you will see that for each motif that was found to be match in the input sequence, information such location of the motif in the sequence, *p*-value, q-value, as well as what the matching sequence is reported.

6. Click “HTML output”. You will be taken a page displaying visual information about the motif scan FIMO performed. Specifically, we can observe that the IDs of the transcription factors that have highly similar motif matches within the upstream sequence of *ABI3*, see **Figure 11**. These are first described by the family of the transcription factor’s DNA binding domain, followed by the AGI ID.



ID of transcription factor motif found within input sequences.

Sequence within input that best matches a given motif.

FIMO - Motif search tool

FIMO version 5.0.5, (Release date: Mon Mar 18 20:12:19 2019 -0700)

For further information on how to interpret these results please access http://meme-suite.org//doc/fimo-output-format.html.
To get a copy of the FIMO software please access http://meme-suite.org

If you use FIMO in your research, please cite the following paper:
Charles E. Grant, Timothy L. Bailey, and William Stafford Noble, "FIMO: Scanning for occurrences of a given motif", *Bioinformatics*, **27**(7):1017-1018, 2011. [full text]

DATABASE AND MOTIFS

```
DATABASE sequences.fa
Database contains 1 sequences, 500 residues

MOTIFS ArabidopsisDAPv1.meme (DNA)

                  MOTIF                 WIDTH  BEST POSSIBLE MATCH
-----------------------------------------------------------------------------
       ABI3VP1_tnt.AT5G18090_col_a_m1     15   GATGATGAAGAAAAG
       ABI3VP1_tnt.AT5G25475_col_a_m1      6   CAAGCA
       ABI3VP1_tnt.AT5G60130_col_a_m1     19   AAGCAAAAAATAAGCAAAA
    ABI3VP1_tnt.AT5G60130_colamp_a_m1     21   TTTTTGCTTATTTTTTGCTTA
            ABI3VP1_tnt.FUS3_col_a_m1     15   ATATACATGCATGCA
            ABI3VP1_tnt.NGA4_col_a_m1     16   CCTGTTCAGGTGATTT
           ABI3VP1_tnt.REM16_col_a_m1     21   TTTACGTTTTTGGCGGGAAAA
            ABI3VP1_tnt.VRN1_col_a_m1     29   TTTTTTTTTTTTTTTTTTTTTTTTTTTTC
         ABI3VP1_tnt.VRN1_colamp_a_m1     19   CTGTTTTTTTTTTTTTTTT
           AP2EREBP_tnt.ABR1_col_a_m1     19   GGCGGAAATGGCGGCGGCG
        AP2EREBP_tnt.ABR1_colamp_a_m1     21   CACCGCCGCCGCCATTTCCGC
           AP2EREBP_tnt.AIL7_col_a_m1     11   CGATTCCCGAG
        AP2EREBP_tnt.AIL7_colamp_a_m1     12   ACGAATCCCGAG
      AP2EREBP_tnt.AT1G01250_col_a_m1     15   GATATTGTCGGTGGT
   AP2EREBP_tnt.AT1G01250_colamp_a_m1     15   TGTCGGTGGAGATGA
      AP2EREBP_tnt.AT1G12630_col_a_m1     21   CGCCGCCGCCTTCGCCGACAA
   AP2EREBP_tnt.AT1G12630_colamp_a_m1     16   GATTTTGTCGGCAAAA
      AP2EREBP_tnt.AT1G28160_col_a_m1     15   TGGCGGCGGCGGTGG
   AP2EREBP_tnt.AT1G28160_colamp_a_m1     14   CGCCGCCGCCGCCA
      AP2EREBP_tnt.AT1G44830_col_a_m1     15   CCACCTCCACCGCCA
      AP2EREBP_tnt.AT1G71450_col_a_m1     20   CACCGCCGCCGCCACCGCCA
```

Figure 11: HTML output of FIMO for the 500-bp upstream region of ABI3. Here, transcription factor motifs that are were found to have the matches with input sequence are displayed.

3.6.2.2 AME- Analysis of Motif Enrichment

Going back to the set of genes that are developmentally coexpressed with *ABI3,* we can query whether these sequences share any putative regulatory similarities. AME can answer this question by finding enriched transcription factor binding sites from a user-supplied motif database.

1. Go to http://meme-suite.org/tools/ame.

2. Under “Select the type of control sequences to use”, toggle “Shuffled input sequences”. Because you chose “Shuffled input sequences”, AME will create control sequences by shuffling the letters in each input sequence.

3. Under “Input the primary sequences”, select “Upload sequences” from the dropdown menu. Upload the file called “ABI3_coexpressed_genes_500bp_upstream.fasta” (available here[16]). This file contains the 500-bp upstream sequences for each of the *ABI3* coexpressed genes in FASTA format.

4. Using the “Input the motifs” drop down menu, we select the database for our search. For the first dropdown menu, select “ARABIDOPSIS (*A. thaliana*) DNA”. Under the second dropdown menu, select “DAP motifs (O’Malley 2016).”

5. Under the “Select the sequence scoring method” drop down menu, select “Average odds score”.

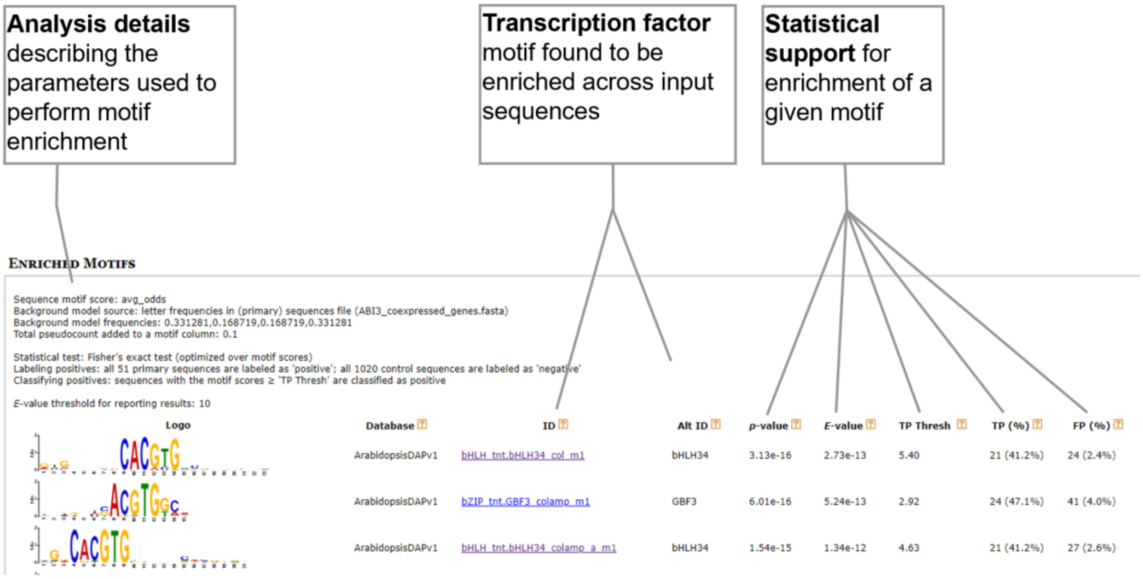
6. Under the “Select the motif enrichment test” drop down menu, select “Fisher’s exact test”.

7. Now, enter in job details, such as your email address and as well as job description describing the AME analysis you are performing. Similar to FIMO, this step will take at least an hour to run on the MEME-suite server, so plan accordingly.

8. When the job is done, you should receive an email with a link to your results. Click the link and you will be taken to an output page that has the following links: AME HTML output, AME TSV output, AME true- and false-positive sequences, and Uploaded Sequences. For future data analysis, download the TSV file. If you open this TSV file, you will see that for each motif that was found to be enriched across the input sequences, information such as the motif consensus sequence, *p*-value, adjusted *p*-value, as well as E-value is supplied. Additional statistical information is also reported, such as “TP thresh”, which the optimal score threshold for determining whether a given sequence is classified as positive for a motif. “TP (%)” describes the percentage of input sequences that are later determined to contain a given motif. Finally, “FP (%)” is the percentage of control sequences (in this example, the shuffled input sequences) that were also found to be matches to a given motif.

[16] Available at http://bar.utoronto.ca/~nprovart/ABI3_coexpressed_genes_500bp_upstream.fasta

9. Click “HTML output”. You will be taken a page displaying visual information about the analysis that you performed with AME. Specifically, you can observe the motif logo plots for their DAP-seq binding sites. The additional information displayed here is the same the TSV that was previously downloaded, see **Figure 12**.



Analysis details describing the parameters used to perform motif enrichment

Transcription factor motif found to be enriched across input sequences

Statistical support for enrichment of a given motif

ENRICHED MOTIFS

Sequence motif score: avg_odds
Background model source: letter frequencies in (primary) sequences file (ABI3_coexpressed_genes.fasta)
Background model frequencies: 0.331281,0.168719,0.168719,0.331281
Total pseudocount added to a motif column: 0.1

Statistical test: Fisher's exact test (optimized over motif scores)
Labeling positives: all 51 primary sequences are labeled as 'positive'; all 1020 control sequences are labeled as 'negative'
Classifying positives: sequences with the motif scores ≥ 'TP Thresh' are classified as positive

E-value threshold for reporting results: 10

Logo	Database	ID	Alt ID	*p*-value	*E*-value	TP Thresh	TP (%)	FP (%)
CACGTG	ArabidopsisDAPv1	bHLH_tnt.bHLH34_col_m1	bHLH34	3.13e-16	2.73e-13	5.40	21 (41.2%)	24 (2.4%)
ACGTGGC	ArabidopsisDAPv1	bZIP_tnt.GBF3_colamp_m1	GBF3	6.01e-16	5.24e-13	2.92	24 (47.1%)	41 (4.0%)
CACGTG	ArabidopsisDAPv1	bHLH_tnt.bHLH34_colamp_a_m1	bHLH34	1.54e-15	1.34e-12	4.63	21 (41.2%)	27 (2.6%)

Figure 12: HTML output of an AME analysis with *ABI3* developmentally coexpressed genes. Here, transcription factor motifs that are enriched in the 500-bp upstream region of the input genes (*i.e.* in their promoters) are ranked by *p*-value.

10. At the top of page, click on “Positive sequences for each motif”. A TSV file containing which input sequences were found to matches to the motifs in the database. We can see that among the top ten most significantly enriched motifs, there are several abscisic acid responsive transcription factors listed: ABF2, ABI5, and AREB3. It would perhaps be interesting to use FIMO to see where these transcription factors bind in the upstream sequences of our coexpressed genes.

3.6.3 ePlant promoter analysis

The ePlant website integrates several essential tools for plant biology research. With only a few mouse clicks the user can examine polymorphisms, visualize gene expression in the whole plant and/or in different tissues, determine the subcellular localization of a protein, find its interactors, and view predicted or experimentally determined protein structures. We will focus on the interactors but let us try a few of the other views first.

1. Go to http://bar.utoronto.ca/eplant.

2. Type the gene name or AGI ID of your gene of interest next into the input box, at the top left of the page. We will use At3g24650 for ABI3.

3. Click on the “Plant eFP” or “Cell eFP” to explore expression levels in the whole plant, in a specific tissue or developmental stage, or to determine where the protein is localized in the cell (in the case of ABI3, mostly in the nucleus). Expression levels are represented from yellow (low) to red (high) in each drawing. Within “Plant eFP Viewers” or “Tissue & Experiment eFP Viewers”, you can use the toolbar at the top to adjust your view. Try toggling between “absolute” and “relative” expression levels – “relative” is useful for perturbation type experiments where one sample serves as a control for others that have been subjected to some sort of stimulus. Additionally, you can click on “Download Raw Data” in the upper grey toolbar to download the numerical gene expression information.

4. Next, click on “Interaction viewer” to view interactors of our gene or gene product. The PDI data used in ePlant comes from the DAP-seq generated in the Ecker Lab (O’Malley et al., 2016) and from Y1H experiments (Brady et al., 2011; Gaudinier et al., 2011; Li et al., 2014; Taylor-Teeples et al., 2015; de Lucas et al., 2016; Murphy et al., 2016; Porco et al., 2016; Sparks et al., 2016), as well as predicted interactions based on FIMO (Grant et al., 2011) mapping. DNA elements are displayed as square and have curved lines indicating interactions with other proteins. As we see in **Figure 13**, ABI3 has a total of 76 protein-DNA interactions, all of which are predicted, as denoted by the grey dashed lines.

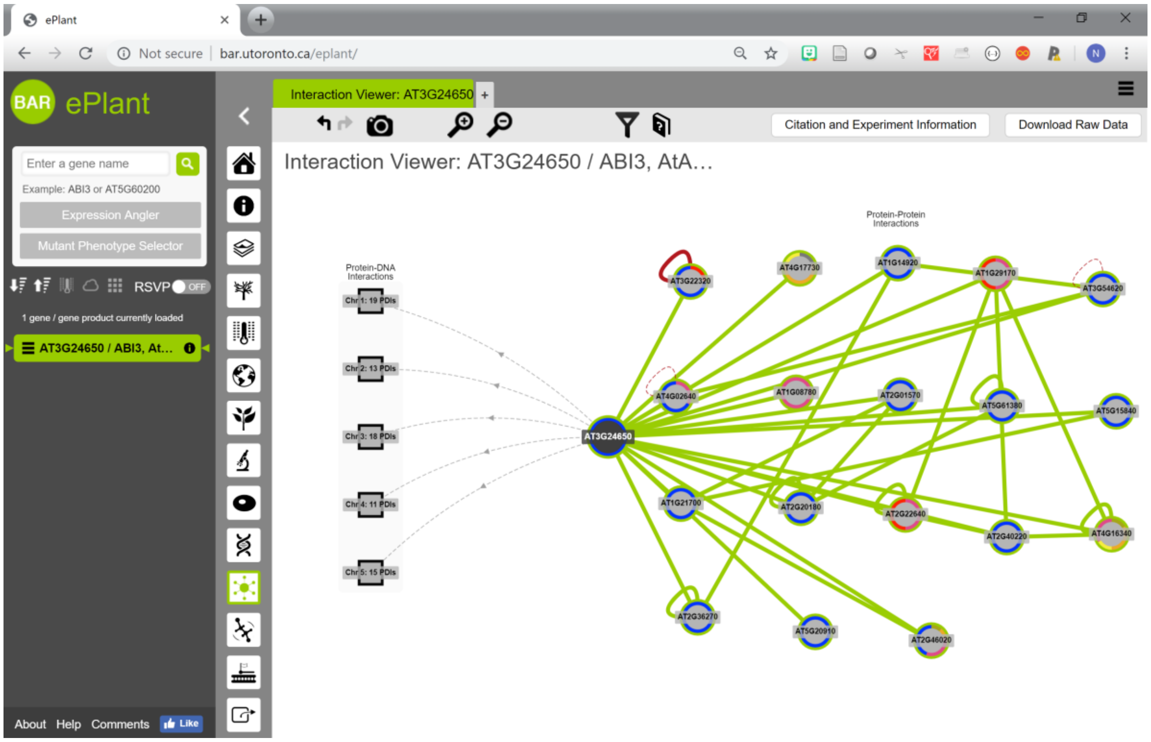


Figure 13: ePlant Interaction Viewer output for ABI3. There are 76 FIMO-predicted DNA interactions for ABI3, as denoted by the grey dashed lines to the 5 chromosome boxes.

3.6.4 TAIR Motif Analysis

Cistome analyzes promoters for over-represented previously validated or characterized regulatory elements. Cistome also provides access to 5 other prediction programs. The Motif Analysis algorithm from TAIR provides an alternate approach by searching for overrepresented 6-mer oligos in upstream regions of genes.

1. Go to http://www.arabidopsis.org/tools/bulk/motiffinder.

2. Add your list of genes by typing the AGI ID or the sequences in Fasta format. Here we will use the list of genes in **Supplementary Table 1**.

3. Indicate the length of the regulatory sequence that will be included in the analysis (*e.g.* 3000bp) and select the output file type (*e.g.* HTML). Click Submit**.**

4. Motif Analysis from TAIR identifies statistically over-represented 6-mer oligos occurring in 3 of more sequences in the gene set. The over-represented 6-mers are sorted by *p*-value determined by comparing against a binomial distribution, and genes/promoters with a particular sequence 6-mer sequence are indicated. One would need to compare these 6-mer sequences to databases already previously described to see if there is overlap with a known sequence. Otherwise, promoter deletion experiments in the lab may show necessity of this sequence for expression, and serve as a starting point for further experimental analyses.

3.7 Functional Classification

Functional classification of gene lists is one of the basic methods in bioinformatics for making sense of sometimes rather large gene lists that arise from gene expression profiling experiments. Typically, one might look at individual genes in such lists and “see if they fit biologically”, but one might also like to have an overview of broad functional categories that change in response to a given stimulus or due to a specific mutation. One of the very useful large initiatives of the past decade was the development of a Gene Ontology (GO) for the “unification of biology” (Ashburner et al., 2000). Basically, this system is based on a set of categories, which are described using defined terms instead of in a free-form manner, into which genes can be assigned. There are three main super-categories: biological process (BP), molecular function (MF), and cellular component (CC). Currently, TAIR is the main curator for GO annotations for *Arabidopsis* genes, with some input from other groups. A gene may belong to several categories and sub-categories at once, which are arranged from very general to very specific terms (technically, the relationships between categories and sub-categories are formalized as a directed acyclic graph).

It is possible to use statistical tests – often a hypergeometric test or Fisher’s exact test– to assess whether the number of genes observed associated with a given term (*i.e.* category) from one’s list of interest is enriched relative to the number one might expect to see by chance. Such tests can be used for any classification system in which objects are classified into categories.

Another system of classification called MapMan Bins was initiated by Björn Usadel and colleagues at the Max Planck Institute for Molecular Plant Physiology in Germany (Thimm et al., 2004). This variation on the approach aims to examine genes whose expression is altered in response to a perturbation in the context of the biological pathways to which they belong.

3.7.1 AgriGO

AgriGO (Tian et al., 2017) out of Zhen Su's laboratory at the Chinese Agricultural University is a user-friendly tool for analyzing whether any particular GO terms are enriched in a given gene list from *Arabidopsis* (or for many other agriculturally-important species). It provides a nice visualization in the same directed acyclic graph structure on which the GO system was developed[17].

1. Go to http://systemsbiology.cau.edu.cn/agriGOv2/ and select "Analysis Tool" in the tab along the top.

2. In the first section for selecting the analysis tool, select "Singular Enrichment Analysis (SEA)".

3. Select the species (the default is *Arabidopsis thaliana*).

4. Paste in the Query list as AGI IDs, gene aliases (*e.g. ABI3*), GenBank IDs etc. A large number of different identifiers are supported.

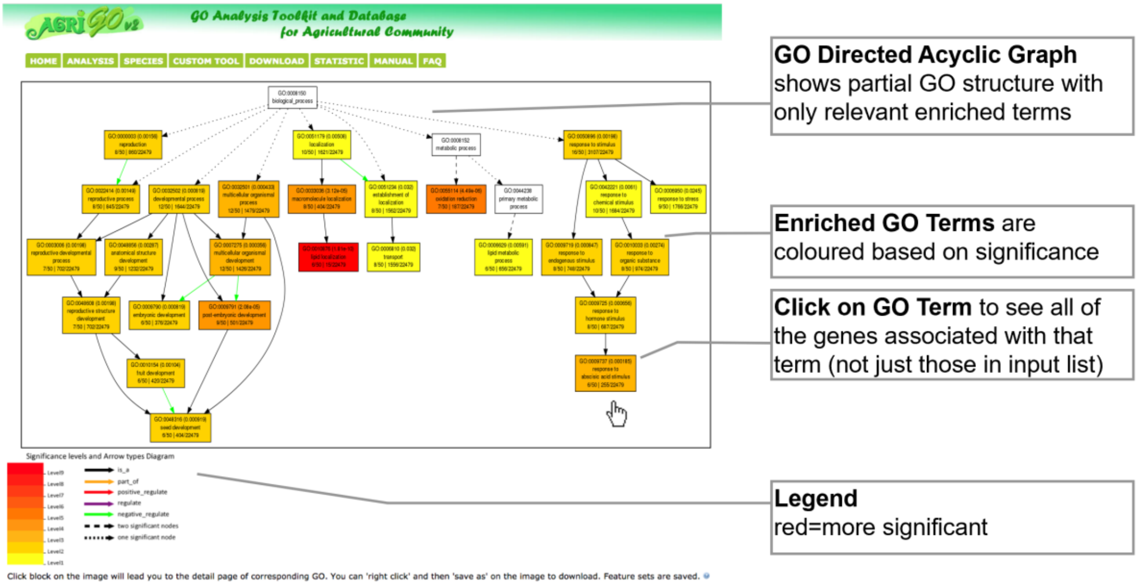
5. Choose a reference – if the list comes from a microarray experiment, then choose the appropriate microarray platform, otherwise if the list comes from an experiment where it is possible to identify *any* of the AGI IDs present in the TAIR genome annotation (such as the case with a proteomics experiment or an mRNA-seq experiment) then choose the "Arabidopsis genome locus (TAIR)" option – this aspect is a nice feature of AgriGO. In this example, we will submit the top 50 genes coexpressed with *ABI3* in the AtGenExpress Tissue Set as discussed in Section 3.5.1, Step 2 (**Supplemental Table 1**). As the data used to obtain the coexpressed genes come from the Affymetrix ATH1 platform, we use this platform as our reference (GPL198).

6. Under "Advanced Options – optional" one can select one of three methods for statistical enrichment (Hypergeometric distribution, Fisher, or Chi-square) as well as one of seven multiple hypothesis testing correction methods. We recommend the use of the Yekutieli method and Fisher's exact distribution (the defaults).

7. In the output, a table of enriched GO categories for our list of 50 genes is displayed showing that four GO Biological Process terms (lipid localization, response to abscisic

[17] GOrilla is another useful tool for such analyses, and permits the ability to upload a ranked list of genes for enrichment analysis. It offers similar visualization of enriched categories. See http://cbl-gorilla.cs.technion.ac.il/.

acid stimulus, macromolecule localization, post-embryonic development) and two GO Molecular Function terms (nutrient reservoir activity, lipid binding) are significantly enriched. Examining these, they seem to “make sense” in the context of the later stages of seed development, when *ABI3* and these genes are expressed, insofar as this is the time when lipid reserves are being accumulated and the seed begins to desiccate. There is also the possibility of creating “Graphical Results” or a “GO Flash Chart”. If we click on the Generate Image button, the following output is generated for enriched Biological Processes (see **Figure 14**).


AgriGO v2
GO Analysis Toolkit and Database for Agricultural Community
HOME | ANALYSIS | SPECIES | CUSTOM TOOL | DOWNLOAD | STATISTIC | MANUAL | FAQ
GO:0010154 (0.00104) fruit development 6/50 | 420/22479
GO:0048316 (0.000919) seed development 6/50 | 404/22479
GO:0009737 (0.000185) response to abscisic acid stimulus 6/50 | 255/22479
Significance levels and Arrow types Diagram
Level9
Level8
Level7
Level6
Level5
Level4
Level3
Level2
Level1
is_a
part_of
positive_regulate
regulate
negative_regulate
two significant nodes
one significant node

GO Directed Acyclic Graph shows partial GO structure with only relevant enriched terms

Enriched GO Terms are coloured based on significance

Click on GO Term to see all of the genes associated with that term (not just those in input list)

Legend red=more significant

Click block on the image will lead you to the detail page of corresponding GO. You can 'right click' and then 'save as' on the image to download. Feature sets are saved.

Figure 14: Graphical output from AgriGO for the top 50 A*BI3*-coexpressed genes in the AtGenExpress Tissue Set from Supplemental Table 1. The GO Biological Process term “Lipid localization” (red) is most significantly enriched among these genes.

3.7.2 AmiGO

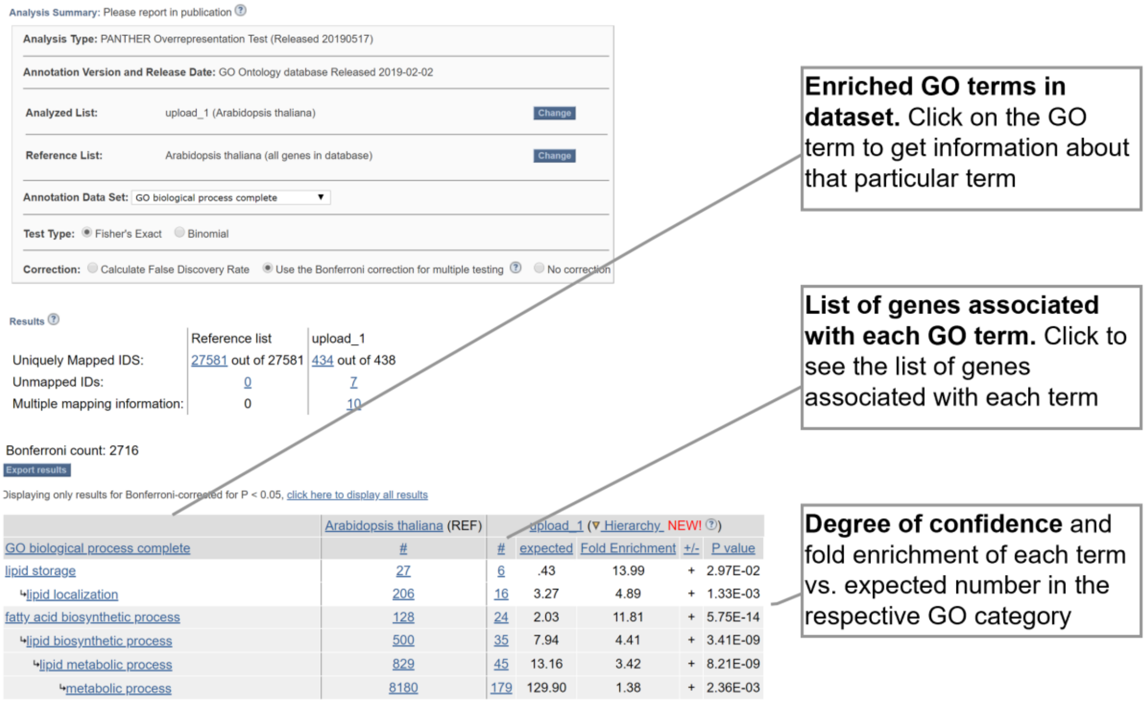
AmiGO (Carbon et al., 2009) provides a generic interface for computing GO term enrichments for all of the species annotated by the GO Consortium.

1. Go to http://amigo.geneontology.org/rte to access the Advanced Options section of the Panther DB enrichment tool.

2. Paste your gene identifiers into the “Gene IDs” box. Using Genevestigator, we have shown that *ABI3* expression depends on the presence of *LEC1.*To better understand this *ABI3-LEC1* relationship, we will use in this instance genes whose steady state transcript level is increased in *LEC1* overexpressor (OX) plants from (Mu et al., 2008), found in Supplemental Table S1 of that article – a text file of these data can be downloaded here… open it with Excel and copy the AGI IDs in the first column. This will allow us to determine biological functions associated with genes that are also overexpressed and likely downstream of *LEC1*. Select *A. thaliana* as the “Species”, and then submit the

query. Click “Submit” (use the Bonferroni correction option). The partial output for this gene list is shown in **Figure 15**.

3. Forty-five genes from the genes up-regulated in LEC1OX plants are grouped into “lipid metabolic process”, with a *p*-value of 8.2e-09 (829 genes in the Arabidopsis genome are associated with this term, meaning an enrichment in our list of ~3.4-fold over the genome-wide background frequency)[18]. *LEC1* is associated to this term, but it is also associated with other several different terms, such as embryo development and others. These analyses indicate that LEC1 is sufficient to regulate lipid metabolism, as was observed for *ABI3* co-expressed genes, and supports that LEC1 likely regulates a module of genes, including *ABI3*, that may coordinate some aspect of lipid metabolism during seed germination.



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (Released 20190517)

Annotation Version and Release Date: GO Ontology database Released 2019-02-02

Analyzed List: upload_1 (Arabidopsis thaliana)

Reference List: Arabidopsis thaliana (all genes in database)

Annotation Data Set: GO biological process complete

Test Type: ● Fisher's Exact ○ Binomial

Correction: ○ Calculate False Discovery Rate ● Use the Bonferroni correction for multiple testing ○ No correction

Results

	Reference list	upload_1
Uniquely Mapped IDS:	27581 out of 27581	434 out of 438
Unmapped IDs:	0	7
Multiple mapping information:	0	10

Bonferroni count: 2716

Displaying only results for Bonferroni-corrected for P < 0.05, click here to display all results

	Arabidopsis thaliana (REF)	upload_1 (▼ Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
lipid storage	27	6	.43	13.99	+	2.97E-02
↳lipid localization	206	16	3.27	4.89	+	1.33E-03
fatty acid biosynthetic process	128	24	2.03	11.81	+	5.75E-14
↳lipid biosynthetic process	500	35	7.94	4.41	+	3.41E-09
↳lipid metabolic process	829	45	13.16	3.42	+	8.21E-09
↳metabolic process	8180	179	129.90	1.38	+	2.36E-03

Enriched GO terms in dataset. Click on the GO term to get information about that particular term

List of genes associated with each GO term. Click to see the list of genes associated with each term

Degree of confidence and fold enrichment of each term vs. expected number in the respective GO category

Figure 15: Output (partial) from AmiGO shows the enriched GO terms in the data set of genes increased in LEC1OX plants, and the genes associated with each GO term (click on # link). Numbers and *p*-values may differ from those shown as GO databases are updated often.

3.7.3 Classification SuperViewer

The BAR’s Classification SuperViewer (Provart and Zhu, 2003) provides a different way to view Gene Ontology and MapMan classifications for lists of genes: it uses a barcode scheme. Classification SuperViewer barcodes are also integrated into several others of the BAR’s output tools.

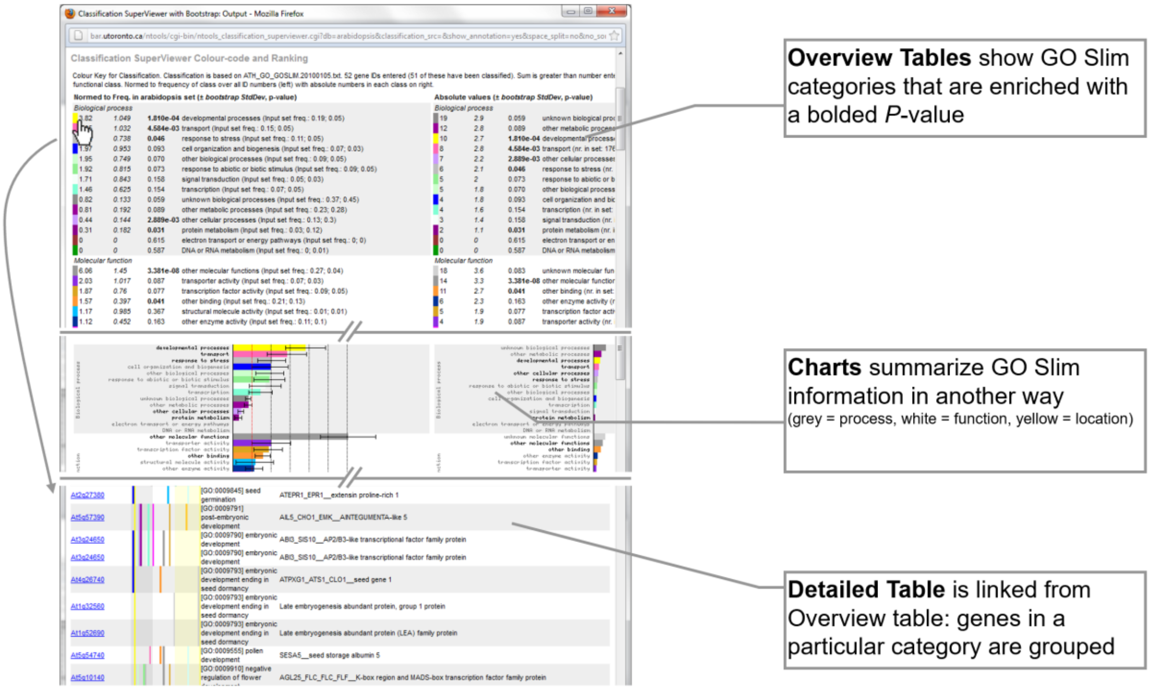
[18] The GO database that AmiGO accesses is updated frequently, and the numbers reported here will likely vary in your own results. These values are for the 2019-02-02 release.

1. Go to http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi and input your list of genes[19].

2. Select the classification scheme you wish to use under the second point, either GO (actually GO Slim in the case of this tool) or MapMan.

3. Leave the other options as they are, and click *Submit Query*.

4. The output page is divided into three parts: an overview table showing which categories are enriched (by a hypergeometric test with a *p*-value cutoff of 0.05) in bold; a chart area summarizing the category information in a different way, and a detailed table section, which is linked from the overview area, see **Figure 16**. In these areas the grey background sections are GO Biological Process terms, those with a white background are GO Molecular Function terms, while those with a yellow background are GO Cellular Component terms (this shading scheme does not apply for MapMan terms).



Classification SuperViewer with Bootstrap: Output - Mozilla Firefox

bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi?db=arabidopsis&classification_src=&show_annotation=yes&space_split=no&no_sor

Classification SuperViewer Colour-code and Ranking

Colour Key for Classification. Classification is based on ATH_GO_GOSLIM.20100105.txt. 52 gene IDs entered (51 of these have been classified). Sum is greater than number ente functional class. Normed to frequency of class over all ID numbers (left) with absolute numbers in each class on right.

Normed to Freq. in arabidopsis set (± *bootstrap StdDev*, p-value)

Biological process			
3.82	*1.049*	**1.810e-04**	developmental processes (Input set freq.: 0.19; 0.05)
[illegible]	*1.032*	**4.584e-03**	transport (Input set freq.: 0.15; 0.05)
[illegible]	*0.738*	**0.046**	response to stress (Input set freq.: 0.11; 0.05)
1.97	*0.953*	0.093	cell organization and biogenesis (Input set freq.: 0.07; 0.03)
1.95	*0.749*	0.070	other biological processes (Input set freq.: 0.09; 0.05)
1.92	*0.815*	0.073	response to abiotic or biotic stimulus (Input set freq.: 0.09; 0.05)
1.71	*0.843*	0.158	signal transduction (Input set freq.: 0.05; 0.03)
1.46	*0.625*	0.154	transcription (Input set freq.: 0.07; 0.05)
0.82	*0.133*	0.059	unknown biological processes (Input set freq.: 0.37; 0.45)
0.81	*0.192*	0.089	other metabolic processes (Input set freq.: 0.23; 0.28)
0.44	*0.144*	**2.889e-03**	other cellular processes (Input set freq.: 0.13; 0.3)
0.31	*0.182*	**0.031**	protein metabolism (Input set freq.: 0.03; 0.12)
0	*0*	0.615	electron transport or energy pathways (Input set freq.: 0; 0)
0	*0*	0.587	DNA or RNA metabolism (Input set freq.: 0; 0.01)
Molecular function			
6.06	*1.45*	**3.381e-08**	other molecular functions (Input set freq.: 0.27; 0.04)
2.03	*1.017*	0.087	transporter activity (Input set freq.: 0.07; 0.03)
1.87	*0.76*	0.077	transcription factor activity (Input set freq.: 0.09; 0.05)
1.57	*0.397*	**0.041**	other binding (Input set freq.: 0.21; 0.13)
1.17	*0.985*	0.367	structural molecule activity (Input set freq.: 0.01; 0.01)
1.12	*0.452*	0.163	other enzyme activity (Input set freq.: 0.11; 0.1)

Absolute values (± *bootstrap StdDev*, p-value)

Biological process			
19	*2.9*	0.059	unknown biological proc
12	*2.8*	0.089	other metabolic proces
10	*2.7*	**1.810e-04**	developmental process
8	*2.8*	**4.584e-03**	transport (nr. in set: 176
7	*2.2*	**2.889e-03**	other cellular processes
6	*2.1*	**0.046**	response to stress (nr.
5	*2*	0.073	response to abiotic or b
5	*1.8*	0.070	other biological process
4	*1.8*	0.093	cell organization and bic
4	*1.6*	0.154	transcription (nr. in set:
3	*1.4*	0.158	signal transduction (nr.
2	*1.1*	**0.031**	protein metabolism (nr. i
0	*0*	0.615	electron transport or en
0	*0*	0.587	DNA or RNA metabolism
Molecular function			
18	*3.6*	0.083	unknown molecular fun
14	*3.3*	**3.381e-08**	other molecular functior
11	*2.7*	**0.041**	other binding (nr. in set:
6	*2.3*	0.163	other enzyme activity (r
5	*1.9*	0.077	transcription factor acti
4	*1.9*	0.087	transporter activity (nr.

Biological process
developmental processes
transport
response to stress
cell organization and biogenesis
other biological processes
response to abiotic or biotic stimulus
signal transduction
transcription
unknown biological processes
other metabolic processes
other cellular processes
protein metabolism
electron transport or energy pathways
DNA or RNA metabolism
other molecular functions
transporter activity
transcription factor activity
other binding
structural molecule activity
other enzyme activity
unknown molecular functions

At2g27380	[GO:0009845] seed germination	ATEPR1_EPR1__extensin proline-rich 1
At5g57390	[GO:0009791] post-embryonic development	AIL5_CHO1_EMK__AINTEGUMENTA-like 5
At3g24650	[GO:0009790] embryonic development	ABI3_SIS10__AP2/B3-like transcriptional factor family protein
At3g24650	[GO:0009790] embryonic development	ABI3_SIS10__AP2/B3-like transcriptional factor family protein
At4g26740	[GO:0009793] embryonic development ending in seed dormancy	ATPXG1_ATS1_CLO1__seed gene 1
At1g32560	[GO:0009793] embryonic development ending in seed dormancy	Late embryogenesis abundant protein, group 1 protein
At1g52690	[GO:0009793] embryonic development ending in seed dormancy	Late embryogenesis abundant protein (LEA) family protein
At5g54740	[GO:0009555] pollen development	SESA5__seed storage albumin 5
At5g10140	[GO:0009910] negative regulation of flower	AGL25_FLC_FLC_FLF__K-box region and MADS-box transcription factor family protein

Overview Tables show GO Slim categories that are enriched with a bolded *P*-value

Charts summarize GO Slim information in another way
(grey = process, white = function, yellow = location)

Detailed Table is linked from Overview table: genes in a particular category are grouped

Figure 16: Output of Classification SuperViewer for a list of the top 50 *ABI3*-coexpressed genes from a query of the BAR's Expression Angler tool in the AtGenExpress Tissue Set compendium (**Supplemental Table 1**). Output values may differ slightly due to database updates.

[19] The Classification SuperViewer is also available for *Medicago truncatula*: http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer_medicago.cgi

5. In the Overview section, categories that are overrepresented relative to the total number of instances of the term in the overall GO or MapMan database[20] are bolded. The relative enrichment is shown on the left, while the absolute number of counts in a given category is on the right. The colour scheme for the categories is also used in the chart section and for the bar code in the table section. In the case of a list of the top 50 genes coexpressed with *ABI3* in the Developmental Map, the Developmental Processes and Transport categories are overrepresented as might be expected for the number of genes in this list involved in the process of dormancy as seeds mature, and in transporting lipids to provide reserves for the seed when it germinates. These categories are also seen with AgriGO.

6. The Chart section shows the overrepresented categories relative to the frequency in the overall Arabidopsis genome or in terms of absolute counts on the left and right side, respectively.

7. The Table sections shows details for every single gene in the input list. A bar code system using the same colour scheme as in the other two sections show that in many cases a given gene falls into several GO categories. Genes are group by category, with the final bar on the right being the category used for grouping. A gene will appear in this table as often as the number of bars in its bar code. Mousing over a particular bar will provide information on the actual GO term.

3.8 Pathway Visualization

One of the biggest challenges working with large-scale data sets is to represent the information in a comprehensive manner. This is particularly relevant in the context of metabolic pathways. If a series of enzymes in a pathway is up-regulated or down-regulated, there is a greater chance that the metabolism of the compounds associated with this pathway will be perturbed accordingly. Pathway visualization tools were generated to integrate and analyze data from large-scale experiments and place that information in an easy-to-interpret metabolic context. In this section we will introduce two different visualization tools used to describe a wide set of *Arabidopsis* metabolic pathways.

3.8.1 AraCyc

AraCyC 8.0 (Mueller et al., 2003) is the most comprehensive *Arabidopsis*-specific metabolic database[21]. We can use their tools to visualize individual metabolic pathways, to view the complete metabolic map of *Arabidopsis*, or to predict metabolic pathways from a list of genes.

[20] Note that it is not possible to select a background data set for Classification SuperViewer. This is not so much of an issue for gene lists that are derived from relatively comprehensive platforms but can be an issue for platforms that are less comprehensive.

[21] AraCyc is a part of the BioCyc metabolic databases. All the metabolic databases present on BioCyc share the same software, so the tutorial described on this section can be applied to the other databases.

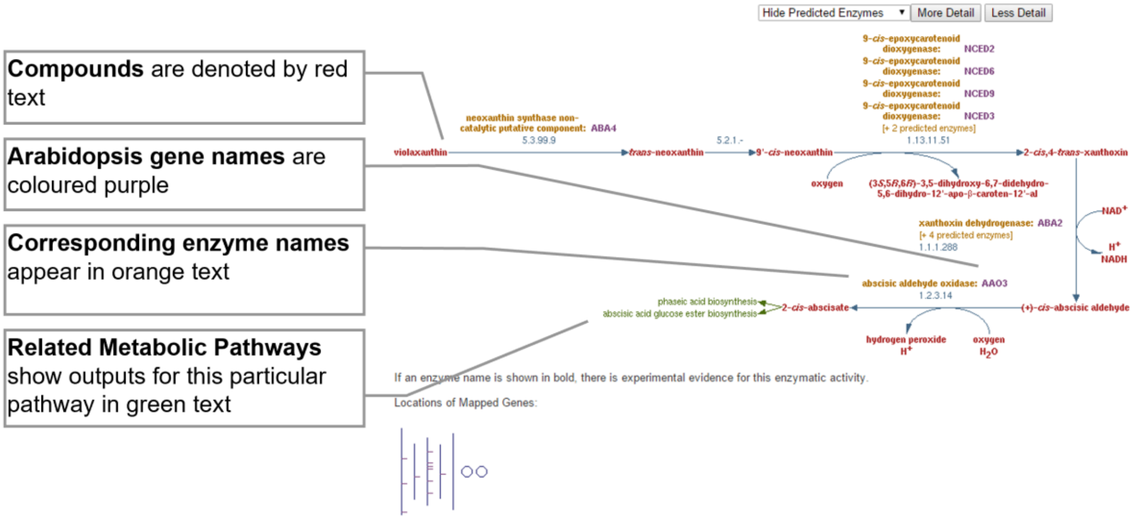
We will demonstrate how to use these three options to characterize the role of ABI3 as it pertains to plant metabolism.

As *ABI3* is highly expressed after treatment with abscisic acid (ABA), we may be interested in learning more about genes that function to synthesize ABA.

1. Go to http://www.plantcyc.org/.

2. In the search box write the name (or a keyword) of the pathway in which you are interested. In our case we will write "Abscisic". Then choose AraCyc as the metabolic database from the dropdown. Click the magnifying glass icon to search.

3. The search results contain a window with a list of pathways, proteins, compounds and reactions that match with our word. We just need to click on the one we want to explore, in our case "abscisic acid biosynthesis" – see **Figure 17**.



Compounds are denoted by red text

Arabidopsis gene names are coloured purple

Corresponding enzyme names appear in orange text

Related Metabolic Pathways show outputs for this particular pathway in green text

Hide Predicted Enzymes ▾
More Detail
Less Detail
9-*cis*-epoxycarotenoid dioxygenase: NCED2
9-*cis*-epoxycarotenoid dioxygenase: NCED6
9-*cis*-epoxycarotenoid dioxygenase: NCED9
9-*cis*-epoxycarotenoid dioxygenase: NCED3
[+ 2 predicted enzymes]
1.13.11.51
neoxanthin synthase non-catalytic putative component: ABA4
5.3.99.9
5.2.1.-
violaxanthin
trans-neoxanthin
9'-*cis*-neoxanthin
2-*cis*,4-*trans*-xanthoxin
oxygen
(3*S*,5*R*,6*R*)-3,5-dihydroxy-6,7-didehydro-5,6-dihydro-12'-apo-β-caroten-12'-al
NAD^+
xanthoxin dehydrogenase: ABA2
[+ 4 predicted enzymes]
1.1.1.288
H^+
NADH
abscisic aldehyde oxidase: AAO3
1.2.3.14
phaseic acid biosynthesis
abscisic acid glucose ester biosynthesis
2-*cis*-abscisate
(+)-*cis*-abscisic aldehyde
hydrogen peroxide
H^+
oxygen
H_2O

If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

Locations of Mapped Genes:

Figure 17: Overview of ABA biosynthesis in AraCyc as resulting from a query with "abscisic acid biosynthesis".

4. AraCyC shows a diagram with the enzymes (orange), compounds (red), genes (purple) and related pathways (green) of the Abscisic Acid biosynthesis pathway. If we click on "more detail" the molecular structures of the compounds appear on the diagram. Below the diagram, we can find information about the chromosomal localization of the genes in the pathway, a brief description of the biological context of the pathway, and the references AraCyc used to generate the pathway.

5. To get information about the enzymatic reaction in which the gene is involved, click on the enzyme name (not the AGI ID). This will take you to a new window with more information. For instance, clicking on the 9-cis-epoxycarotenoid dioxygenase will give all interactions in which this enzyme could be involved in, as well as the enzymatic reactions of all closely related homologs.

6. To get detailed information on the gene, double-click on the gene name. For instance, *ABA4* (*At1g67080*) encodes a neoxanthin synthase involved in the conversion of violaxathin into trans-neoxanthin, which is an early step in ABA biosynthesis. We can expect that mutants in *ABA4* have reduced levels of ABA, hence the expression of *ABI3* will be reduced too since it is ABA-responsive (Brady et al., 2003). Transcriptome analysis of *aba4* mutants would be useful to study the plant's behaviour in the absence of ABA to determine any correlation with loss of *ABI3* function.

In section 3.4.4 (Genevestigator) we determined that *ABI3* was up-regulated in *LEC1* over-expression plants (pER8-LEC1). We will use the list of genes up-regulated in LEX1-OX plants (Mu et al., 2008) to predict metabolic pathways that *LEC1* overexpression modulated with the OMICs Viewer tool of the AraCyC database. Download the file here. These genes may act with *ABI3* to influence plant form or function.

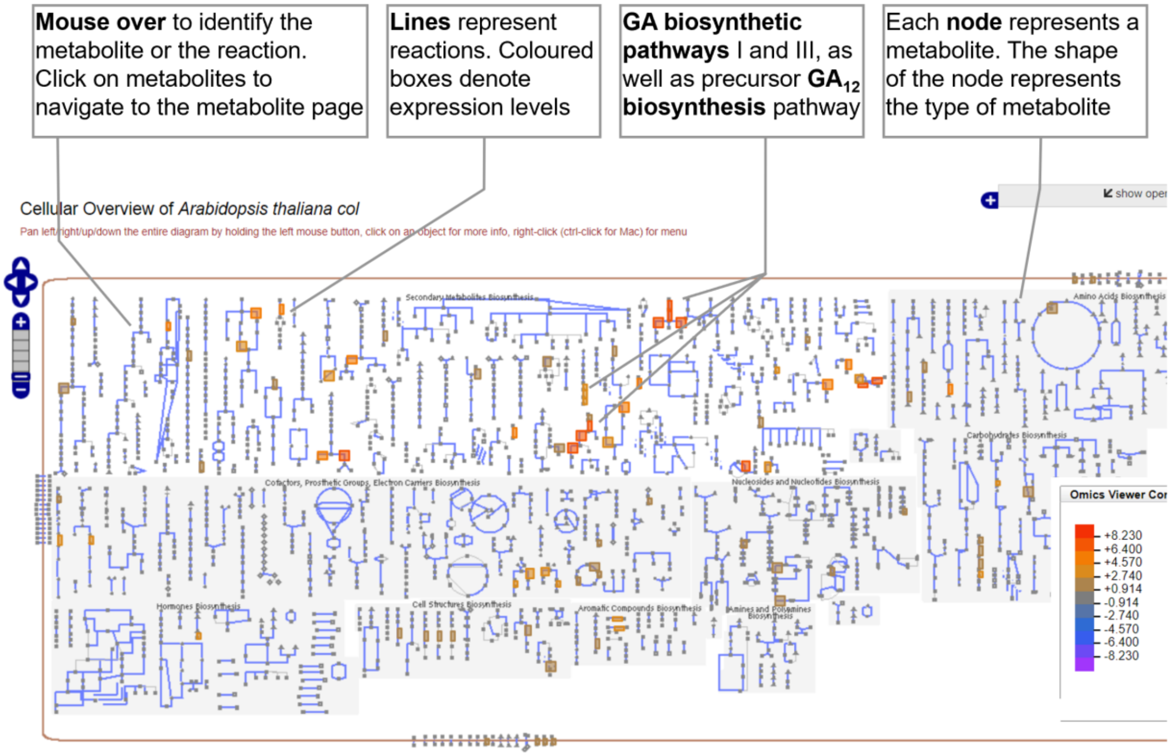
1. Go to http://pmn.plantcyc.org/overviewsWeb/celOv.shtml (you can access this from the Metabolism tab at the top of the AraCyc page, too).

2. In the right part of the window, the OMICs Viewer presents various operations we can perform with it. Click on the "Upload Data from File" option. The file must be in tab-delimited text format and the first column must be the locus name (*e.g.* *At3g24650*) and the second the expression value[22]. Click on "Browse" to upload the file. Choose "Relative" or "Absolute" values to display. We have only one column of relative expression data. As our data are $\log_2$-transformed, we will use the "0-centred scale". We are using locus names in our data, so choose "Gene names and/or identifiers" as the items that appear in the first column of our data file. In our data file we only have one experiment, so type "1" in the *Data columns to use* box (if your data has multiple set of values, type the numbers of the columns you want to display). We can also play with colour scheme options and display type. We will leave the other options as their defaults. Click "Submit".

3. The output window shown in **Figure 18** shows a diagram with all metabolic pathways of *Arabidopsis*. The OMICs viewer uses red to represent highly expressed genes. Multiple genes involved in gibberellin biosynthesis appear to be highly up-regulated and over-represented in our expression data which suggests that GA biosynthesis may be up-regulated in the *LEC1* overexpression line.

4. To see in detail the pathways represented in our expression data, go back to the "Upload Data from File" part of the OMICs viewer and Show Data "As a table of pathway diagrams" and select the number of pathways to show, such as 25 (the OMICS Viewer uses a "Pathway Perturbation Score" to rank the pathways according to their overall increase or decrease in expression. The pathways are shown in a table.

[22] We can include more expression columns: each one could represent a different experiment or time point.



Mouse over to identify the metabolite or the reaction. Click on metabolites to navigate to the metabolite page

Lines represent reactions. Coloured boxes denote expression levels

GA biosynthetic pathways I and III, as well as precursor **GA$_{12}$ biosynthesis** pathway

Each **node** represents a metabolite. The shape of the node represents the type of metabolite

Cellular Overview of *Arabidopsis thaliana col*
Pan left/right/up/down the entire diagram by holding the left mouse button, click on an object for more info, right-click (ctrl-click for Mac) for menu
show oper
Secondary Metabolites Biosynthesis
Amino Acids Biosynthesis
Carbohydrates Biosynthesis
Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis
Nucleosides and Nucleotides Biosynthesis
Hormones Biosynthesis
Cell Structures Biosynthesis
Aromatic Compounds Biosynthesis
Amines and Polyamines Biosynthesis
Omics Viewer Cor
+8.230
+6.400
+4.570
+2.740
+0.914
-0.914
-2.740
-4.570
-6.400
-8.230

Figure 18: Partial output of AraCyc's OMICs viewer summarizing the increases in transcript abundance in LEC1 overexpression plants.

5. LEC1OX appears to promote gibberellin biosynthesis though the activation of genes involved in that metabolic pathway, such as GA20 oxidases 3 and 7. LEC1 acts as a positive regulator upstream of ABI3 (Baud et al., 2002), as *ABI3* is up-regulated in LEC1 plants. As we have seen in data sets contained in Genevestigator, the GA biosynthetic inhibitor paclobutrazol inhibits *ABI3* expression. It appears that LEC1 and ABI3 could play a role in the crosstalk between ABA and GA pathways, which supports the known influence of these genes in these pathways.

3.8.2 Pathway Visualization – MapMan

One of the most widely used software for pathway visualization is MapMan (Thimm et al., 2004). This software classifies genes and metabolites in ontologies based on metabolic pathway, cellular function, biological response and gene families. The main advantage is that the user can download the software and work offline. Also, the databases associated with MapMan are well annotated and are easily downloadable in a format that is useful for bioinformaticians.

1. Go to http://mapman.gabipd.org/web/guest/mapman-download and download the latest version of MapMan[23]. Open MapMan.

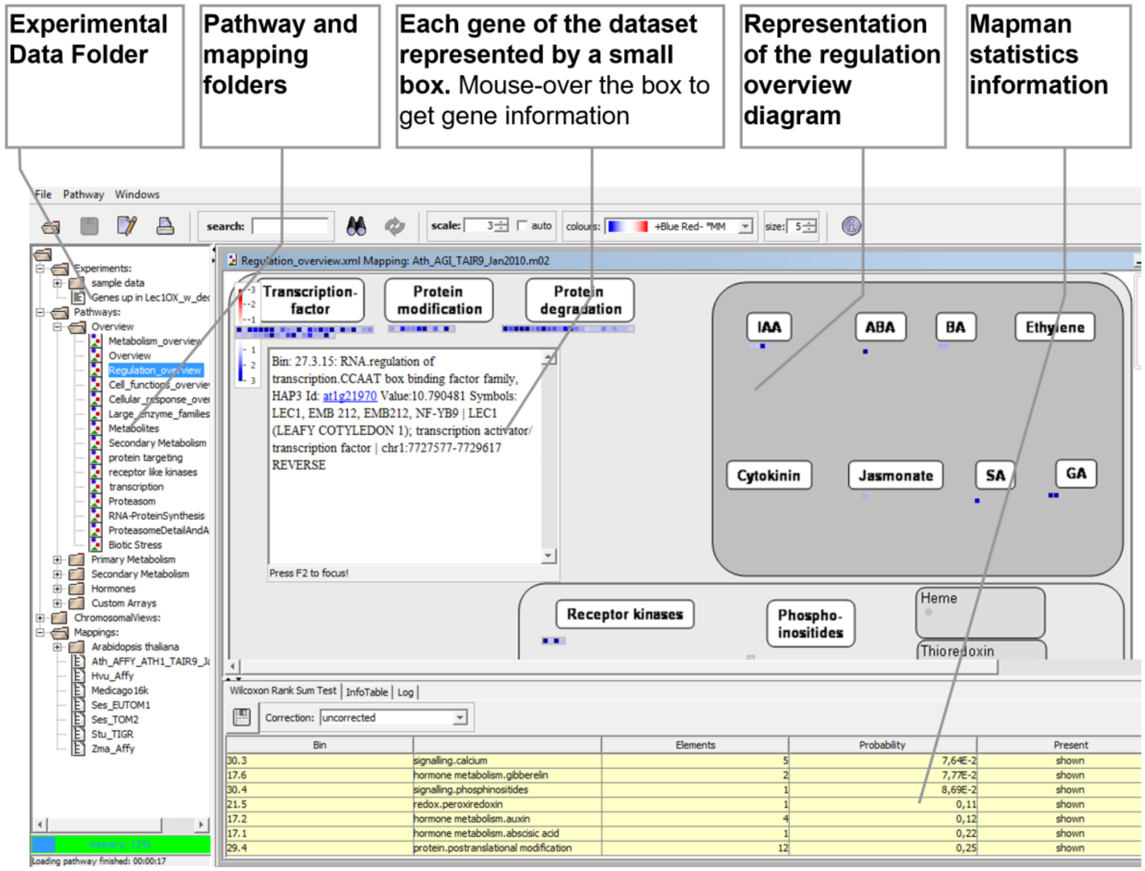
[23] The MapMan version used in this chapter is 3.6.0RC1.

2. Once open, the software shows the “get started” window that will help us on the tool use. Basically, MapMan works by combining a data file (experimental results) with diagrams (pathways or chromosomal views) and mapping information. Every file is stored in a specific folder (left side of the program). Before starting the analysis, it is worth exploring the files available in MapMan (pathways and mapping files). To download more pathways or mapping files from the MapManStore server, click “File”, “Add pathway” or “Add mapping”, click “Download” and choose a pathway/map from the list, *i.e.* download the last gene TAIR annotation.

3. Upload your data by clicking on the folder icon at the top left of the screen. Data must be in .xls or a tab delimited .txt file, the first column should contain the AGI ID (or Affymetrix ID) numbers and the second column, the expression values. The data will be stored in the “Experiments” folder. We will use the genes up-regulated in the LEC1OX plants present in Mu et al. (2008). Download the file here if you didn’t already for the AraCyc OMICS viewer section.

4. For visualization of the data, choose a pathway from the left and double click, *i.e.* “Regulation overview”. Choose a mapping according to the data. If the data contain AGI IDs use Ath_AGI_TAIR (if they contain Affymetrix IDs, use Ath_AFFY_TAIR). For LEC1OX genes, click on the data file uploaded in step 3.

5. MapMan shows a representation of the pathways and genes showing altered regulation, see **Figure 19**. Each gene is symbolized by a square and expression is colour encoded (by default red denotes down-regulated, blue denotes up-regulated). As we are looking at over-expressed genes in the LEC1OX, we only see blue colours. We can see that LEC1 overexpression promotes the expression of transcription factors, genes involved in protein modification and degradation. Looking at hormone pathways, we can see that LEC1 promotes the expression of genes involved in auxin, brassinosteroid and gibberellin metabolism. Below the pathway representation, there is information about the statistical enrichment (using the Wilcoxon rank sum test) performed in MapMan. Mouse over gene squares to see information about gene function, name and expression value. More information about how to use MapMan with experimental data is provided in an online tutorial on the MapMan site.



Experimental Data Folder

Pathway and mapping folders

Each gene of the dataset represented by a small box. Mouse-over the box to get gene information

Representation of the regulation overview diagram

Mapman statistics information

Bin: 27.3.15: RNA.regulation of transcription.CCAAT box binding factor family, HAP3 Id: at1g21970 Value:10.790481 Symbols: LEC1, EMB 212, EMB212, NF-YB9 | LEC1 (LEAFY COTYLEDON 1); transcription activator/ transcription factor | chr1:7727577-7729617 REVERSE

Press F2 to focus!

Bin		Elements	Probability	Present
30.3	signalling.calcium	5	7,64E-2	shown
17.6	hormone metabolism.gibberelin	2	7,77E-2	shown
30.4	signalling.phosphinositides	1	8,69E-2	shown
21.5	redox.peroxiredoxin	1	0,11	shown
17.2	hormone metabolism.auxin	4	0,12	shown
17.1	hormone metabolism.abscisic acid	1	0,22	shown
29.4	protein.postranslational modification	12	0,25	shown

Figure 19: Output of a MapMan pathway analysis using genes upregulated in LEC1OX plants (see text).

3.9 Protein Information

3.9.1 Protein Information – SUBA4

The subcellular location database for *Arabidopsis* proteins is a comprehensive resource encompassing experimental ("direct assay") data from more than 1,768 publications, in which 11,740 proteins have annotated subcellular localizations based on chimeric fusion studies or subcellular proteomic studies (Hooper et al., 2017). In addition, subcellular localization predictions generated by 22 algorithms are also provided. It is possible to specify which you would like to retrieve from the SUBA database on the input page. Alternatively, one can query in a general manner, either for a single gene or for a list of genes, as follows:

1. Go to http://suba.live/. Click on the "Search" tab, then the "Protein name & properties" tab and scroll down to the section that is labeled "alias, family IDs or protein sequence". Type in "ABI3". Click on the green "+" icon to add it to the search criteria, then scroll up and click on the button labeled "Query".

2. You will be taken to a new tab with the query results.

3. Click on “AT3G24650.1” under AGI. You will be redirected to a database factsheet page.

4. We know ABI3 is a transcription factor, but if we look at the Predictors window, users can also see that see that ABI3 can be located in the mitochondria and endoplasmic reticulum, amongst other cellular compartments.

5. Go back to the first page (http://suba.live) and click the “Clear” button beside the “Query” button. Assume we know that we want to investigate a mysterious transcription factor related to FUS3 and LEC1. Under “protein description” (not where you entered the gene name before), type in FUS3, then click on the green “+” button. Delete the contents of the text box and replace it with “LEC1” and “transcription factor”, clicking on the green “+” button in between.

6. Click on the query button. This should return ABI3.

3.9.2 Protein Information – Cell eFP Browser

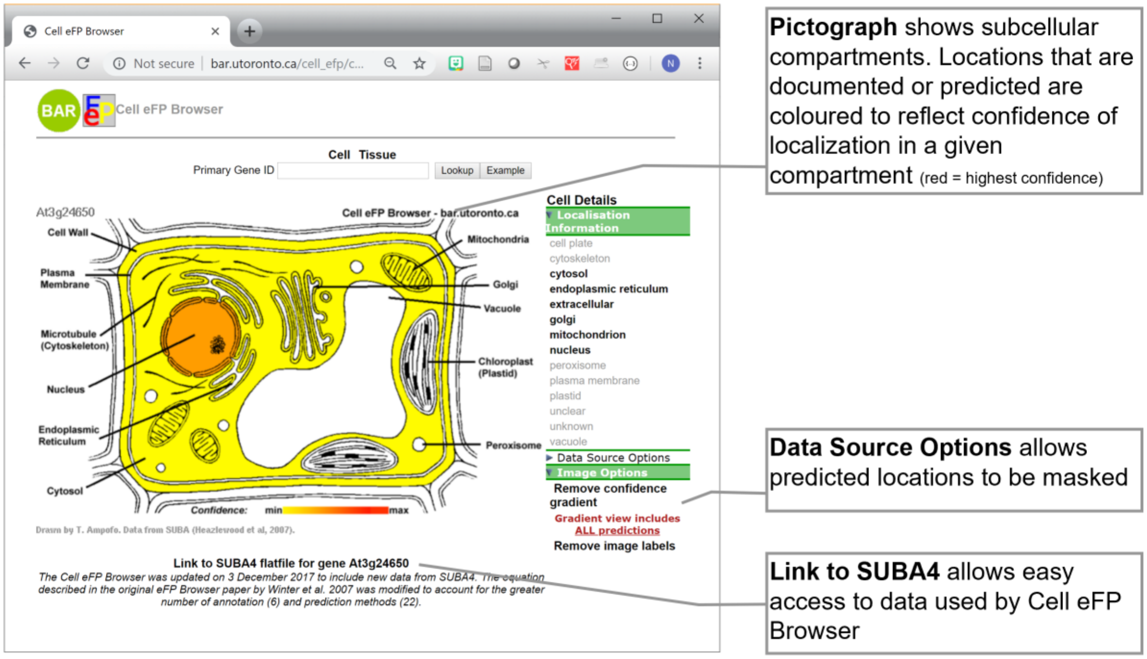
There is an alternative, easier way to obtain general information on subcellular localizations. The Cell eFP Browser (one of the many plant bioinformatic tools available at the Bio-Analytic Resource for Plant Biology) is a simple tool that generates a graphical view of the predicted and experimentally-determined localizations generated/curated by SUBA (Winter et al., 2007). Cell eFP uses a simple heuristic algorithm that weighs “direct assay” subcellular localization data higher than prediction programs to provide a visual representation of where the protein is localized within the cell. It provides a similar view to the Cell eFP view in ePlant (Section 3.6.3). The colour scale is not dynamically adjusted, however.

1. Go to http://bar.utoronto.ca/cell_efp/cgi-bin/cell_efp.cgi.

2. Enter the AGI ID for a gene of interest, for example At3g24650. Click “Lookup”.

3. On the output page a pictograph will be displayed showing the localization of the protein, see **Figure 20**. A stronger red colour denotes that several direct assays have documented the protein being at a particular location. Predictions receive a weighting only one fifth of that for direct assays.

4. It is possible to adjust the data sources used for display by using the boxes on the right side of the Cell eFP output.



The Cell eFP Browser was updated on 3 December 2017 to include new data from SUBA4. The equation described in the original eFP Browser paper by Winter et al. 2007 was modified to account for the greater number of annotation (6) and prediction methods (22).

Pictograph shows subcellular compartments. Locations that are documented or predicted are coloured to reflect confidence of localization in a given compartment (red = highest confidence)

Data Source Options allows predicted locations to be masked

Link to SUBA4 allows easy access to data used by Cell eFP Browser

Figure 20: Cell eFP Browser output for ABI3. Colouring points to a nuclear localization for ABI3.

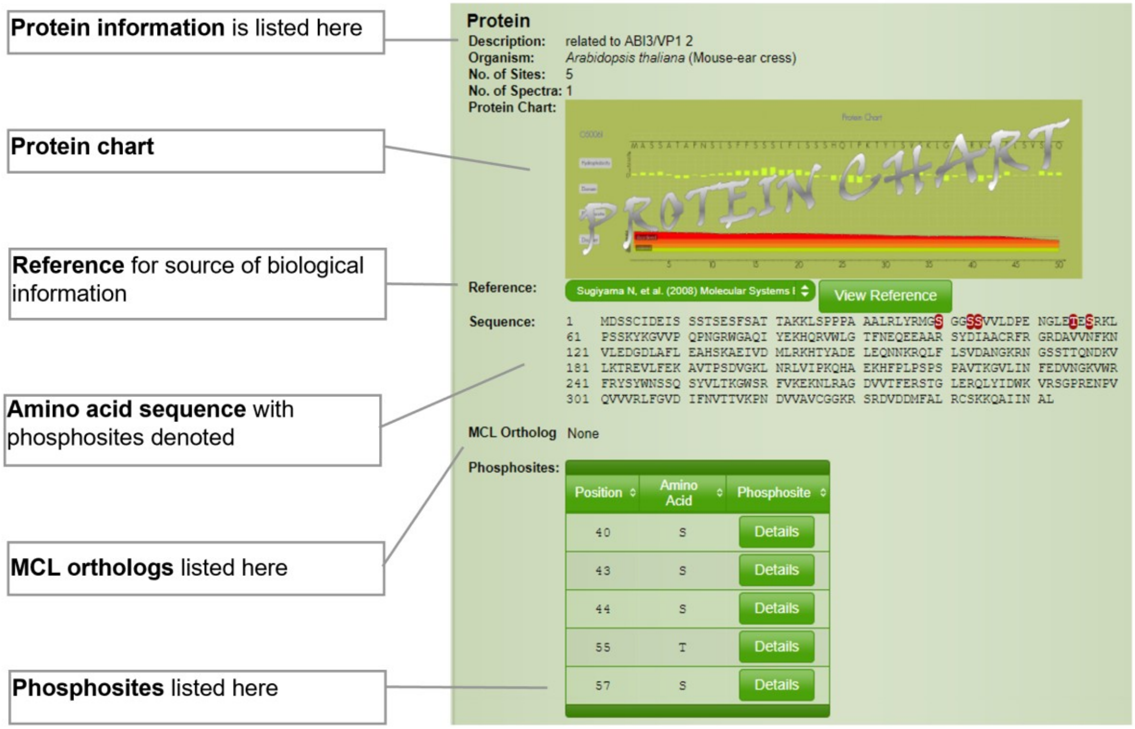
3.9.3 P^{3}DB – Plant Protein Phosphorylation Database

P^{3}DB is a plant protein phosphorylation and acetylation database that contains widespread *in vivo* phosphoproteomic data for many plant species, including *A. thaliana* (Gao et al., 2009; Yao et al., 2014, 2012). As of 2013, P^{3}DB contains curated data describing ~50,000 phosphosites and ~16,000 phosphoproteins across 9 plant species. Data included in the database are protein-protein interactions (PPI), Gene Ontology, protein tertiary structures, orthologous sequences, kinase/phosphatase classification, as well as Kinase Client (KiC) Assay data. It is an excellent place to start if you are uncertain about whether your protein contains phosphorylated sites.

1. Go to http://p3db.org/index.php. Under “Quick Search” at the top of the page, type in “ABI3”.

2. You will be taken to a new page showing four entries for protein, protein description, and in which species the protein is found.

3. Click on “Details” for *A. thaliana*. You will be taken to new a page that displays experimental details about ABI3 in *A. thaliana*. For instance, you can see under the “Sequence” section of the page that there are three serines and two threonines that were found to be phosphorylated (**Figure 21**). If we click on any of these amino acids, we will be taken to new page displaying additional experimental details about how this peptide fragment was identified. If we did not know that phosphorylation of our protein of interest was important for its regulation, as is the case for ABI3 (Yang et al., 2017), we could hypothesize that post-translational modification is important for its regulation.


Protein information is listed here
Protein chart
Reference for source of biological information
Amino acid sequence with phosphosites denoted
MCL orthologs listed here
Phosphosites listed here
Protein Chart
PROTEIN CHART
5
10
15
20
25
30
35
40
45
50

Protein

Description: related to ABI3/VP1 2
Organism: *Arabidopsis thaliana* (Mouse-ear cress)
No. of Sites: 5
No. of Spectra: 1
Protein Chart:

Reference: Sugiyama N, et al. (2008) Molecular Systems I View Reference

Sequence:

```
1   MDSSCIDEIS SSTSESFSAT TAKKLSPPPA AALRLYRMGS GGSSVVLDPE NGLETESRKL
61  PSSKYKGVVP QPNGRWGAQI YEKHQRVWLG TFNEQEEAAR SYDIAACRFR GRDAVVNFKN
121 VLEDGDLAFL EAHSKAEIVD MLRKHTYADE LEQNNKRQLF LSVDANGKRN GSSTTQNDKV
181 LKTREVLFEK AVTPSDVGKL NRLVIPKQHA EKHFPLPSPS PAVTKGVLIN FEDVNGKVWR
241 FRYSYWNSSQ SYVLTKGWSR FVKEKNLRAG DVVTFERSTG LERQLYIDWK VRSGPRENPV
301 QVVVRLFGVD IFNVTTVKPN DVVAVCGGKR SRDVDDMFAL RCSKKQAIIN AL
```

MCL Ortholog None

Phosphosites:

Position	Amino Acid	Phosphosite
40	S	Details
43	S	Details
44	S	Details
55	T	Details
57	S	Details

Figure 21: Portion of P^{3}DB output for ABI3.

3.9.4 Plant PTM Viewer

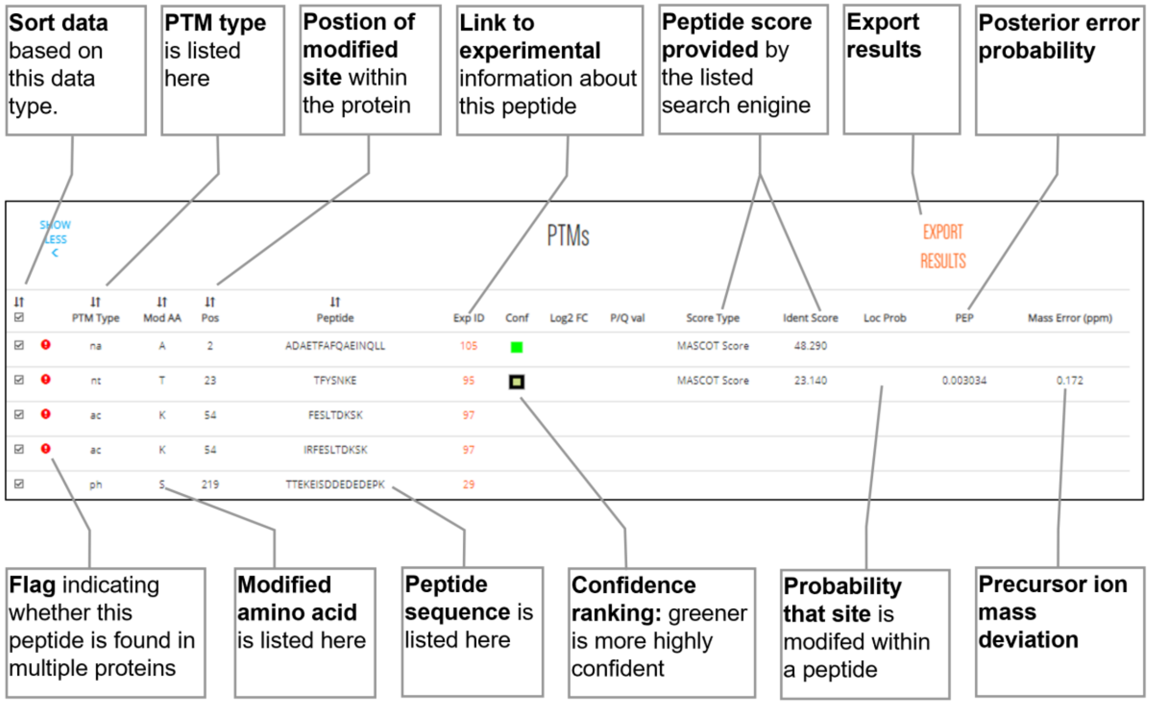
Similar to P^{3}DB, Plant PTM Viewer is an integrative, centralized database of proteomics-detected post-translational modifications (PTMs), such as phosphorylation sites, in many plant proteins. Unlike P^{3}DB, Plant PTM Viewer contains data describing other PTMs such as, but not limited to, methylation, nitrosylation, ubiquitination, and glycosylation (Willems et al., 2019). In total, 37,000 PTM sites are reported for five different plant species, including Arabidopsis. Plant PTM Viewer is also an open repository of published data sets, therefore the limiting factor on data availability is what has been published. Plant PTM Viewer also contains detailed and comprehensive tutorials for site utility, which can be found here: https://dev.bits.vib.be/ptm-viewer/tutorial.php.

1. Go to https://dev.bits.vib.be/ptm-viewer/index.php. Click on “Protein Search” at the top of the page.

2. Instead of ABI3, which does not currently have any protein modifications listed in this repository, let’s search for HSP90.1 (HEAT SHOCK PROTEIN 90.1) using the AGI ID AT5G52640.

3. At the bottom of the page your search results should be displayed. Click on the ID “AT5G52640.1”.

4. You will now be taken to a new page displaying the experimental data describing PTMs of HSP90.1. As you can see, there are eleven amino acid sites that span six different PTMs. If you scroll down a bit, you can also view the eleven amino acids that are modified in their sequence context. You may also notice that the “Conf” column, there are green squares. Confidence scores are based on extracted peptide scores from a variety of sources and Plant PTM viewer labels PTMs as either low, medium or high confidence (Willems et al., 2019). The greener the square, the more experimental validations are associated with this PTM. For details on this calculation, please see Willems et al. (2019). See **Figure 22** for additional explanations of the output columns.



⇅ ☑		⇅ PTM Type	⇅ Mod AA	⇅ Pos	⇅ Peptide	Exp ID	Conf	Log2 FC	P/Q val	Score Type	Ident Score	Loc Prob	PEP	Mass Error (ppm)
☑	❗	na	A	2	ADAETFAFQAEINQLL	105	🟩			MASCOT Score	48.290			
☑	❗	nt	T	23	TFYSNKE	95	▣			MASCOT Score	23.140		0.003034	0.172
☑	❗	ac	K	54	FESLTDKSK	97								
☑	❗	ac	K	54	IRFESLTDKSK	97								
☑		ph	S	219	TTEKEISDDEDEDEPK	29								

Sort data based on this data type.

PTM type is listed here

Postion of modified site within the protein

Link to experimental information about this peptide

Peptide score provided by the listed search enigine

Export results

Posterior error probability

Flag indicating whether this peptide is found in multiple proteins

Modified amino acid is listed here

Peptide sequence is listed here

Confidence ranking: greener is more highly confident

Probability that site is modifed within a peptide

Precursor ion mass deviation

Figure 22: Image showing expanded results from the Plant PTM Viewer for HSP90.1.

5. You can also observe that there are several amino acids with red exclamation points next to them in the “PTM type” column. This flag indicates that the peptide with this PTM is also found in other proteins. If you hover over the red exclamation points, the genes that share this peptide will be listed. If you search these other genes, you will find that they are HSP90 paralogs. Mechanistically, this should make sense because paralogs typically have highly similar genic sequences.

6. To export the data, click on “Export results” at the top of the page. A TSV file containing information about PTMs of HSP90.1 should be downloaded automatically.

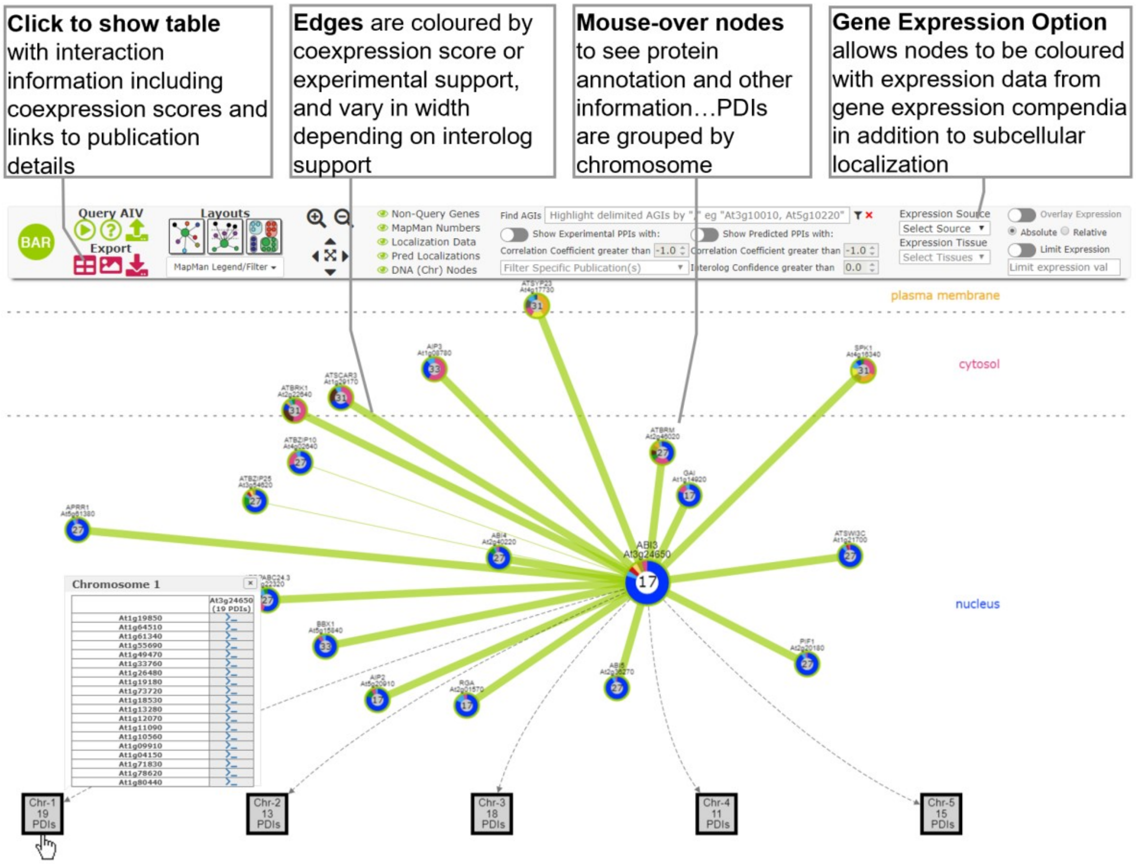
3.10 Protein-Protein Interaction Networks

There are several databases to explore for *Arabidopsis* protein-protein interactions. A big Arabidopsis-specific one is the BAR's new Arabidopsis Interactions Viewer 2 (AIV2). However, it is advisable to examine other databases, such as IntAct at http://www.ebi.ac.uk/intact/ (Orchard et al., 2014) or BioGRID at http://thebiogrid.org (Chatr-Aryamontri et al., 2017), both of which are not specific for *Arabidopsis*, or AtPID (http://www.megabionet.org/atpid/) by Li et al. (2011), as literature curation efforts are by no means complete for any of these databases. Note that as of a couple of years ago, the AIV2 described in the next section is able to automatically query PSICQUIC-enabled databases for other Arabidopsis interactions, thereby facilitating the searching of multiple databases (PSICQUIC is an effort to standardize the access to molecular interaction databases, see https://github.com/PSICQUIC/).

3.10.1 Arabidopsis Interactions Viewer 2 (AIV2)

The BAR's Arabidopsis Interactions Viewer 2 at http://bar.utoronto.ca/interactions2/ (Dong et al., 2019) currently permits the exploration of 80,009 predicted and 62,626 experimentally-determined protein-protein interactions curated by BIND, the BAR, IntAct, TAIR etc., along with ~2.8M protein-DNA interactions (PDIs). One may submit a list of gene (product) identifiers and the AIV will return the interactors of the proteins. It is possible to return only experimentally documented interactions, or all interactions including those predicted through the use of the interolog method (*inter*acting orth*olog*) described in Geisler-Lee et al. (Geisler-Lee et al., 2007) or via docking (Dong et al., 2019). Attractive features of the AIV include the ability to upload Cytoscape files (.cys files) as well as the ability to colour nodes by their expression level in different tissues to help define subnetworks in different tissue types. Unlike ePlant, a nice "layered" view from the outside of the cell to the inside of the cell is available, and, for PDIs, a matrix of interactions is available to help ascertain which DNA targets have transcription factors binding in common.

1. Go to http://bar.utoronto.ca/interactions2/.
2. Enter an AGI identifier, or a list of identifiers, and select any of the options you wish. The default setting will return all experimentally determined and predicted interactions for your gene products of interest. For this example we will not check any of the additional options, and we will again use ABI3, At3g24650 to search for proteins with which it interacts. You can select to search from multiple databases that store Arabidopsis interactions by ticking the appropriate boxes below the AGI ID input field.
3. Click "Submit".
4. On the output page, a network graph of ABI3 interactors appears, plus a legend, some further options, and a table of these interactors at the bottom of the page, see **Figure 23**.



Click to show table with interaction information including coexpression scores and links to publication details

Edges are coloured by coexpression score or experimental support, and vary in width depending on interolog support

Mouse-over nodes to see protein annotation and other information…PDIs are grouped by chromosome

Gene Expression Option allows nodes to be coloured with expression data from gene expression compendia in addition to subcellular localization

Find AGIs: Highlight delimited AGIs by "," eg "At3g10010, At5g10220"

Show Experimental PPIs with: Correlation Coefficient greater than -1.0

Filter Specific Publication(s)

Show Predicted PPIs with: Correlation Coefficient greater than -1.0

Interolog Confidence greater than 0.0

Expression Source: Select Source

Expression Tissue: Select Tissues

Overlay Expression ● Absolute ○ Relative

Limit Expression: Limit expression val

Chromosome 1

	At3g24650 (19 PDIs)
At1g19850	≻_
At1g64510	≻_
At1g61340	≻_
At1g55690	≻_
At1g49470	≻_
At1g33760	≻_
At1g26480	≻_
At1g19180	≻_
At1g73720	≻_
At1g18530	≻_
At1g13280	≻_
At1g12070	≻_
At1g11090	≻_
At1g10560	≻_
At1g09910	≻_
At1g04150	≻_
At1g71830	≻_
At1g78620	≻_
At1g80440	≻_

Figure 23: Output page of an Arabidopsis Interactions Viewer 2 query with At3g24650, ABI3.

5. In the network graph, the smaller nodes represent the proteins that interact with ABI3, and the edges denote the interactions between the proteins. Node colour indicates protein subcellular localization, partitioned according to support for a given localization. Edges coloured in green indicate interactions for which experimental evidence was obtained.

6. The default output is for the nodes to be coloured according to their subcellular localization as documented in the SUBA4 database (see above). A useful feature is to colour nodes according to their expression levels in a given tissue. Clicking the *Expression Source / Select Source* dropdown on the top right of the output screen. After you have chosen the Expression Source, you can then *Select Tissues*. The "Source" option allows you to explore different compendia (the same ones as visible in the various eFP Browser views described earlier), while the "Tissues" option allows you to choose which tissue or condition within a given compendium you are interested in using to retrieve expression level data for painting onto the nodes. In this case, we will examine the expression levels in *Seeds Stage 10 w/o Siliques* in the *Developmental_Map* data source by selecting these and toggling the *Overlay Expression* button. These data are mostly from Schmid et al. (Schmid et al., 2005). In this case we see that ABI3 and ABI5 (but not the other interactors) are both strongly expressed in the seeds at later stages of development, consistent with their known biological roles. It is possible to explore the

expression levels for the corresponding nodes (genes) by selecting different data sets and tissues/conditions to permit you to identify other tissues in which other nodes are more strongly expressed (*e.g. Tissue_specific* / *Guard Cells no ABA*).

3.11 Integrated Tools

Integrated tools associate data from multiple heterogeneous sources of genomic data to obtain more accurate predictions. Most of the bioinformatics tools described in this section integrate protein and genetic interactions, pathways, coexpression, co-localization and protein domain similarity and allow the user generate hypotheses in a rapid and facile manner.

3.11.1 VirtualPlant

VirtualPlant (Katari et al., 2010) integrates genomic data from different sources[24] and provides a set of tools to visualize and analyze these data. One extremely useful attribute of VirtualPlant is that data and analyses can be stored on the website.

1. Go to http://virtualplant.bio.nyu.edu/cgi-bin/vpweb/. If you wish to store your data, click on “Login” to register. The dark blue navigation bar at the top of the page contains the different VirtualPlant tools.

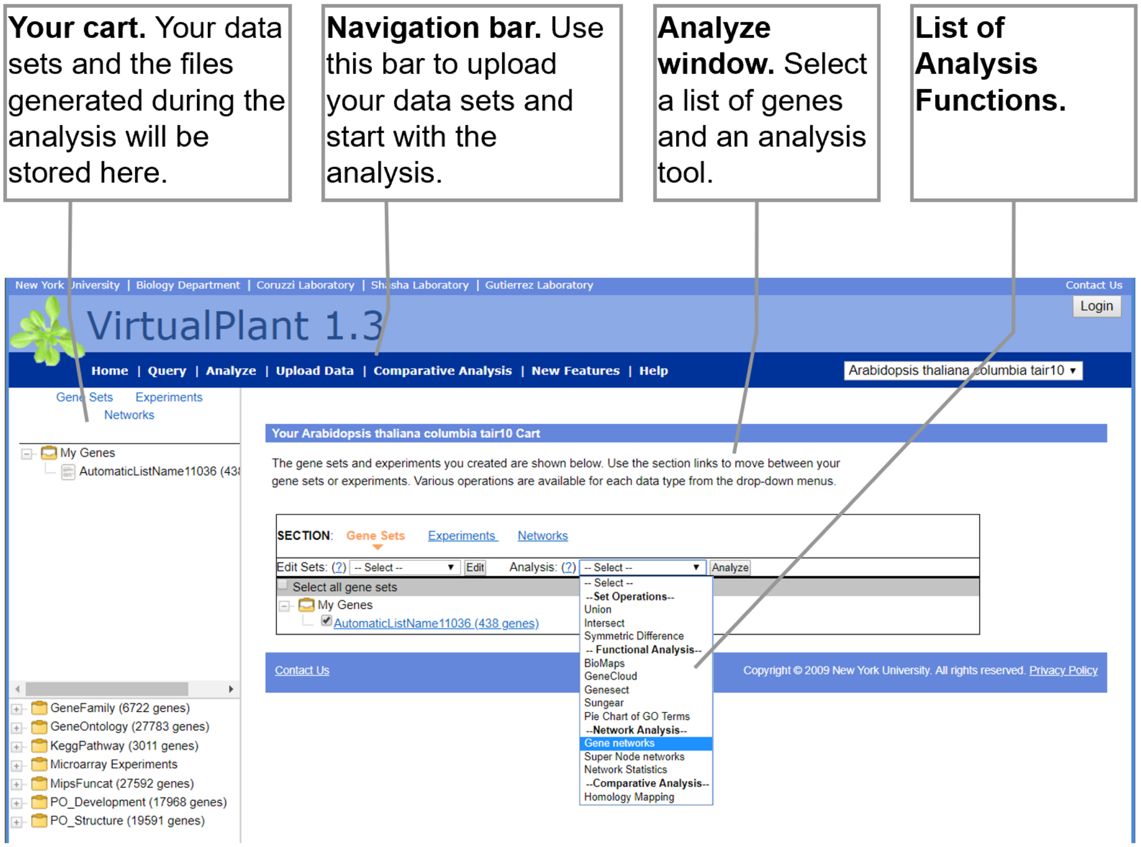
2. Click on “Query”. To perform a query, select an option on the type list (*i.e.* genes) and add a keyword (*e.g.* ABI3). The results are displayed in a table; click on the gene that best matches your query (*i.e. ABI3*, At3g24650). VirtualPlant shows all the information available on the server about our query, including annotation, gene models, and external links. For additional data click on the “Gene Family” folder to see more members of the ABI3VP1 transcription factor family - the ABI3VP1 family has 11 members.

3. To analyze a list of genes, data must be uploaded. The user can upload a list of genes or microarray experiments. One useful feature of VirtualPlant is that for microarray analysis, .CEL files can be uploaded and normalized (GCRMA or MAS5 methods) using VirtualPlant.

4. In the dark blue toolbar click on “Upload Data” followed by “Click here to upload one or more list of genes” and paste your list of genes or upload a file following the format described at the top of the page, or paste your list of genes, *i.e.* paste the AGI IDs from the list of genes upregulated in LEC1OX plants (Mu et al., 2008). Download the file here if you didn’t already for the AraCyc OMICS viewer section and open it with Excel to copy the first column of AGI IDs. Click “Submit”. Our list of genes is now uploaded to the “My Genes” folder. Click “Analyze” in the navigation bar. The analysis window shows the gene sets in our folder – select our data set (see **Figure 24**). In the “Analyze” menu, select the experiment you want to perform. One of the most useful analysis tools available on VirtualPlant is the “Network Analysis” tool. Here, with a list that is available one can select from a variety of interactions including validated TF-target, microRNA-

[24] VirtualPlant currently integrates information from Arabidopsis and rice sources.

target mRNA, and metabolic and pathway interactions from KEGG and AraCyc. An independent Cytoscape browser (“Virtual Plant meets Cytoscape”) is launched – see **Figure 25** (again, you may need to add http://virtualplant.bio.nyu.edu to the list of permitted sites in your Java security settings). One can explore the different interactions by colouring the edges with different colours in Cytoscape via the VizMapper tool. In this case we can determine that the majority of genes overexpressed within LEC1 are metabolic in nature.



Your cart. Your data sets and the files generated during the analysis will be stored here.

Navigation bar. Use this bar to upload your data sets and start with the analysis.

Analyze window. Select a list of genes and an analysis tool.

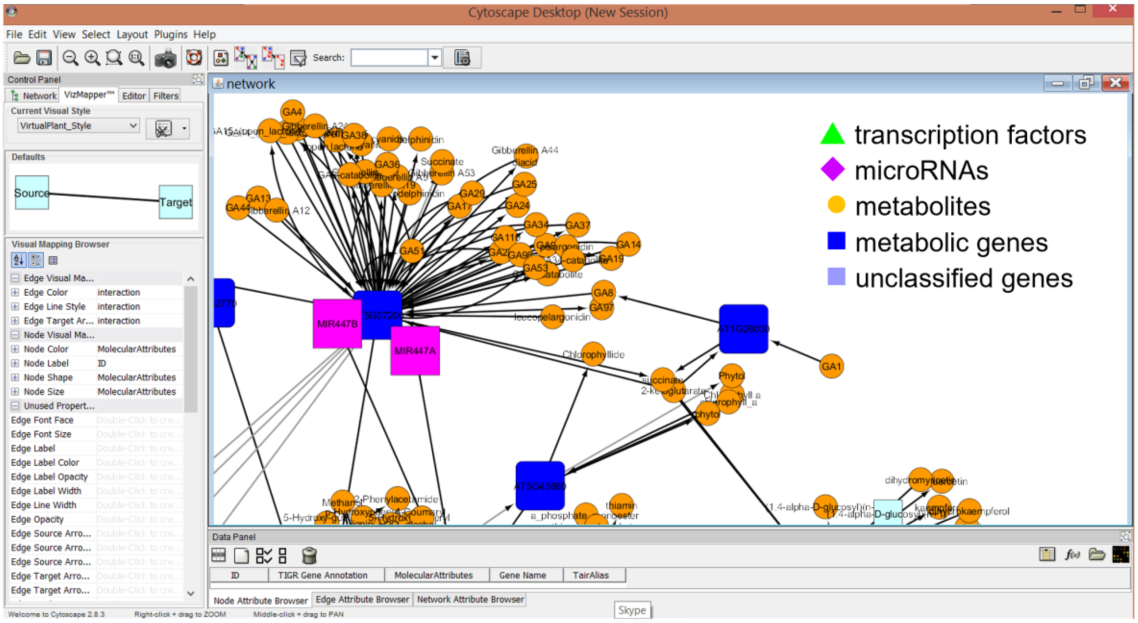
List of Analysis Functions.

Your Arabidopsis thaliana columbia tair10 Cart

The gene sets and experiments you created are shown below. Use the section links to move between your gene sets or experiments. Various operations are available for each data type from the drop-down menus.

Copyright © 2009 New York University. All rights reserved. Privacy Policy

Figure 24: VirtualPlant workspace.



Current Visual Style: VirtualPlant_Style

Edge Visual Ma...	
Edge Color	interaction
Edge Line Style	interaction
Edge Target Ar...	interaction
Node Visual Ma...	
Node Color	MolecularAttributes
Node Label	ID
Node Shape	MolecularAttributes
Node Size	MolecularAttributes
Unused Propert...	
Edge Font Face	Double-Click to cre...
Edge Font Size	Double-Click to cre...
Edge Label	Double-Click to cre...
Edge Label Color	Double-Click to cre...
Edge Label Opacity	Double-Click to cre...
Edge Label Width	Double-Click to cre...
Edge Line Width	Double-Click to cre...
Edge Opacity	Double-Click to cre...
Edge Source Arro...	Double-Click to cre...
Edge Source Arro...	Double-Click to cre...
Edge Source Arro...	Double-Click to cre...
Edge Target Arro...	Double-Click to cre...
Edge Target Arro...	Double-Click to cre...

ID	TIGR Gene Annotation	MolecularAttributes	Gene Name	TairAlias

Welcome to Cytoscape 2.8.3

Figure 25: A snapshot of the Cytoscape graph output from VirtualPlant. A gibberellic acid metabolism-associated gene can be seen at the top left (square blue node connected to many circular orange nodes). These associations come from KEGG or AraCyc, but the metabolic genes were flagged as such from our input list of LEC1OX up-regulated genes. A few microRNAs (square magenta nodes) target a couple of unclassified genes, not shown in this view.

5. VirtualPlant also allows the analysis of multiple gene lists at the same time. We may be interested in finding common genes between the two experiments. We could, for instance, determine if there are any genes that are upregulated when *LEC1* is overexpressed and that are coexpressed with *ABI3*. This would identify that LEC1 is sufficient to regulate these genes, which also may share functionality with ABI3.

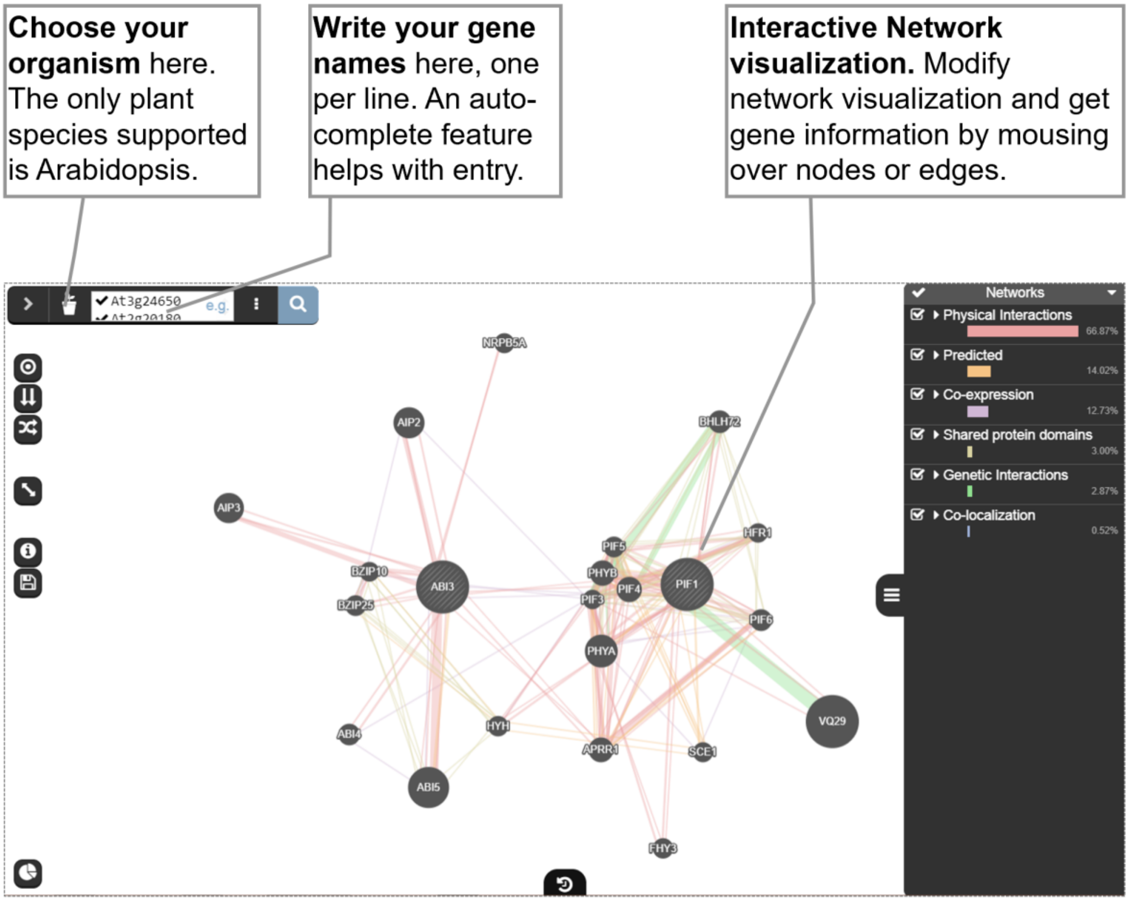
3.11.2 GeneMania

The GeneMania (Mostafavi et al., 2008) algorithm uses a Cytoscape plugin to integrate protein and genetic interaction data, coexpression and co-localization information. We can use GeneMania to predict the function of a single gene or to find new members of a pathway or a protein complex. In the steps below, we will explore the relationship between PIF1 and ABI3.

1. Go to http://www.genemania.org/.
2. GeneMania integrates data from seven different organisms. In the top left select the *Arabidopsis thaliana* icon and add your gene or list of genes. GeneMania recognizes gene names and AGI IDs. If GeneMania doesn’t recognizes your query it will tell you with a yellow speech bubble. We will add ABI3 (At3g24650) and PIF1 (At2g20180 – also called PIL5) in the second window (one gene per line) to try to predict a mechanism for why *ABI3* expression is downregulated in *pif1pif3pif4pif5* quadruple mutant plants.

3. On the left part of the window a network graph visualized using Cytoscape is displayed with coloured edges to indicate different interaction types between different genes. Brown indicates predicted interactions, grey indicates coexpression, dark blue indicates physical interactions, light blue indicates co-localization and green indicates genetic interactions. On the right side of the window, there are four tabs. The “network” tab gives us the option to select the type of interactions we want to see on the right diagram, *e.g.* we can check the physical interactions tab only. It appears that PIL5 could form a protein complex with ABI3 and At5g61380. There are many examples by which protein complexes can have autoregulatory function on one or more of the members of the protein complex (Cui et al., 2007). By clicking on the nodes that represent the genes, we get more details of a gene’s function. For instance, At5g61380 is a two-component response regulator and possesses transcription regulatory activity. The “gene” tab gives a list of interactors with our query proteins, *e.g.* the DELLA protein interacts with PIL5/PIF1. It has been described that DELLAs repress PIF activity and that they are accumulated in the absence of GA (de Lucas et al., 2008; Dill et al., 2001). This could potentially be the mechanism by which negative crosstalk exists between ABA and GA. The “Functions” tab shows the GO annotation of the genes in the network. We can sort the list by GO annotation name by the False Discovery Rate of by Coverage (number of genes in the network with a given function divided by all the genes in the genome with that function), see **Figure 26**.

4. Above the network diagram there is a bar with more options to save the data or to play with network graph visualization.



Choose your organism here. The only plant species supported is Arabidopsis.

Write your gene names here, one per line. An auto-complete feature helps with entry.

Interactive Network visualization. Modify network visualization and get gene information by mousing over nodes or edges.

Figure 26: GeneMania output, with Coexpression and Physical interaction networks shown.

3.11.3 ePlant

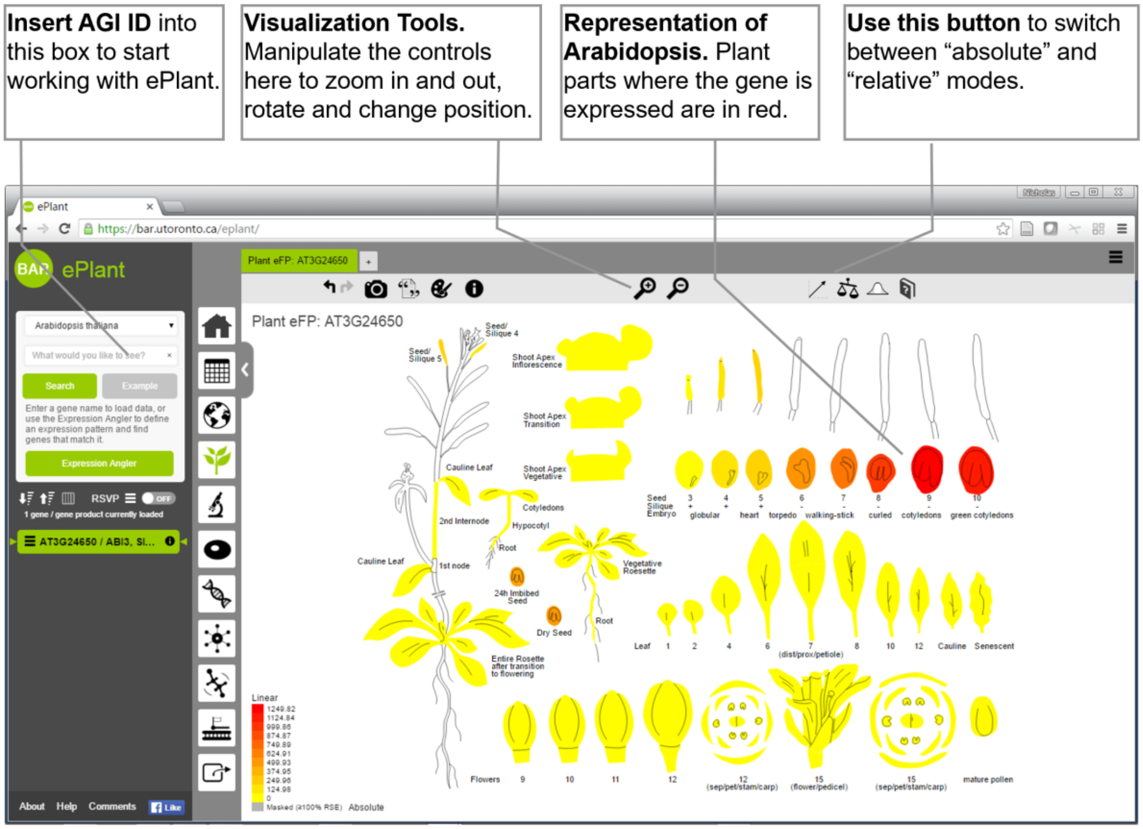
The ePlant website (Waese et al., 2017) integrates several essential tools for plant biology research. With only a few mouse clicks the user can examine polymorphisms, visualize gene expression in the whole plant and/or in different tissues, determine the subcellular localization of a protein, find its interactors, and view predicted or experimentally determined protein structures.

1. Go to http://bar.utoronto.ca/eplant.

2. Type the gene name or AGI ID of your gene of interest next into the input box, at the top left of the page, *e.g.* At3g24650 for *ABI3.*

3. Click on “World eFP”, “Plant eFP” or “Cell eFP” to explore expression levels in the whole plant, in a specific tissue or developmental stage or to determine where the protein is localized into the cell. For each analysis, ePlant uses SVG graphics that represents the *Arabidopsis* plant, different plant tissues or a plant cell (**Figure 27**). Expression levels

are represented from yellow (low) to red (high) in each drawing[25]. On the left part of each page, there are tools to manipulate the visualization. The user can zoom in and zoom out. On the "Plant expression" and "Tissue expression", under "sample list" buttons to change from "absolute" to "relative" expression levels. Click on "Retrieve signal data" to get and/or download the numerical gene expression information. ABI3 is highly expressed on seed siliques, and it is not expressed in the root, leaves, stem or flowers. At the developmental and tissue-specific level, ABI3 is expressed in dry and imbibed seeds. With the "Cell eFP" tool, ABI3 is depicted with high confidence (based on both experimentally determined and computationally predicted results) as being localized within the nucleus.



Insert AGI ID into this box to start working with ePlant.

Visualization Tools. Manipulate the controls here to zoom in and out, rotate and change position.

Representation of Arabidopsis. Plant parts where the gene is expressed are in red.

Use this button to switch between "absolute" and "relative" modes.

Figure 27: ePlant "Plant eFP" output for *ABI3*.

4. Click on "Interactor Viewer" to view interactors with our gene. ePlant uses the BAR's AIV2 database to generate and graph a network with edges and nodes. On the top of the page, there is a funnel icon to adjust the network properties, *i.e.* we can filter the interactors according to the confidence value of the edges (CV). Click on a node and "Get Data" to load that protein/gene into ePlant.

[25] In the case of subcellular location, information comes from SUBA4 database. The shading denotes the protein localization, with red representing a more likely compartment.

5. Click on “Molecule Viewer” to view a 3D structure of your protein. ePlant shows a 3D model from the Protein Data Bank or predicted by Phyre2[26] (Protein HomologY/analogy Recognition Engine) using JSmol. The options on the right of the page allow the user to highlight the molecular surface. Below the structure, ePlant represents the alignment between the sequence used for the 3D model and the query protein, *e.g.* the ABI3 3D model represent amino acids 566 to 678 of the protein. Right click on the model for other visualization options. You can also map Pfam domains and CDD Feature hits onto the 3D structure by selecting the appropriate button (if available for your protein of interest), or identify any non-synonymous (amino acid-changing) polymorphisms from the 1001 Genomes database, which are visualized as pins, sized and coloured according to the frequency of the polymorphism.

3.11.4 TF2Network

We saw with ePlant and AIV2 that it is possible to start to think about networks of genes controlling the expression of other genes, perhaps combinatorially. With TF2Network (Kulkarni et al., 2018) we can start with a list of coexpressed genes and then see if there are potentially common regulators of those genes. TF2Network uses position weight matrices and scores the occurrence of each PWM in the coexpressed gene promoters using hypergeometric tests. It assigns *p*-values, corrected by the Benjamini-Hochberg method. The top 50 TFs, according to *p*-value, are visually reported as predicted regulators. Experimentally determined PPIs and PDIs are also incorporated into the output.

1. Go to: http://bioinformatics.psb.ugent.be/webtools/TF2Network/.

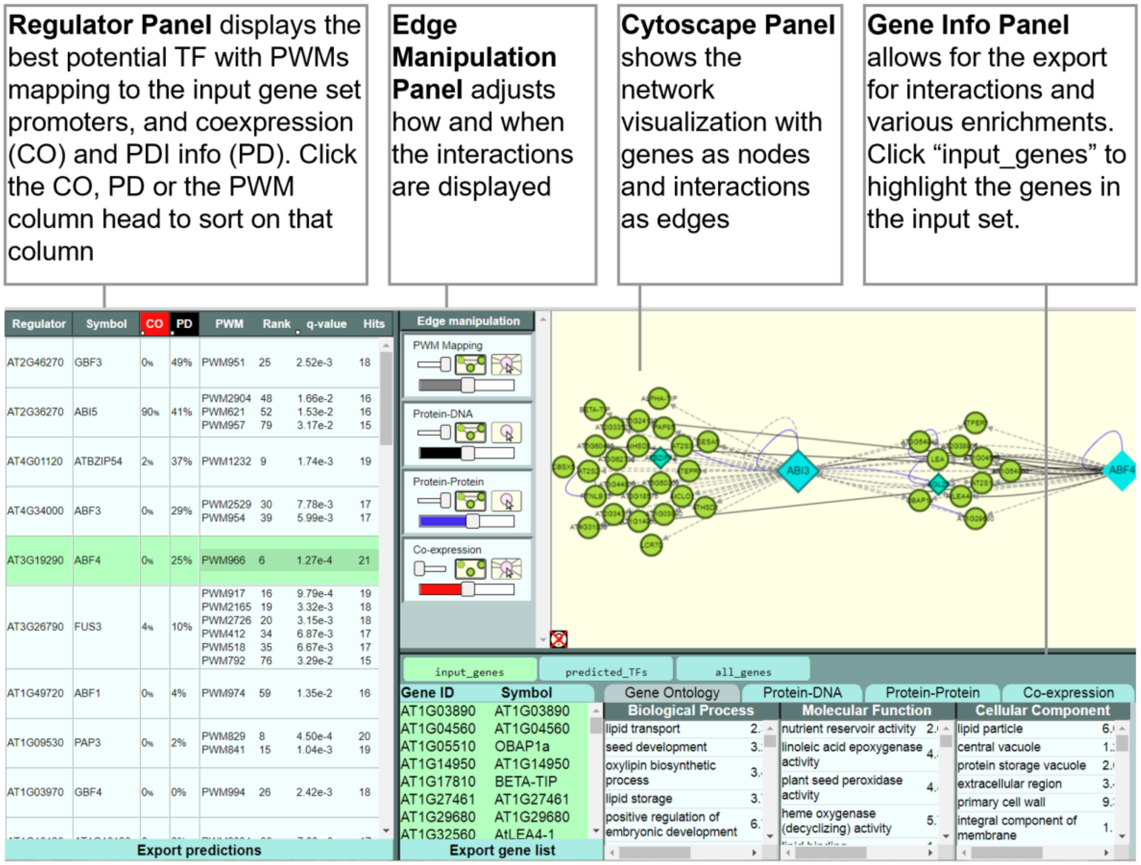
2. Let’s use the top 50 coexpressed genes for ABI3 across a “Developmental Map” as identified with the Expression Angler tool (Supplemental Table 1). We will copy the AGI IDs and paste them into the into the input box. Then click “submit”.

3. Select the ABI3 TF by clicking on its row in the left panel of the output to view its interactions in the Cytoscape panel on the right. Manipulate the edges to explore its different interactions by using the smaller sliders in the “Edge Manipulation panel”. Use the mouse wheel to zoom into the network. Click on node for this gene and examine the information in the gene info panel.

4. Select an additional TF in the regulator panel (*e.g.* ABF4/At3g19290) to see interactions between the two genes and the input gene set. It is possible to sort by CO (coexpression column indicating what percent of the input genes are coexpressed – in a TF2Network internal coexpression database – with the inferred regulator), PD (experimentally-determined protein-DNA data column showing percent of input genes having PDI data) or the PWM column (the number of PWM matches to input set promoters is shown in the

[26] Protein Data Bank: http://www.rcsb.org/pdb/home/home.do. Phyre website: http://www.sbg.bio.ic.ac.uk/phyre2/.

Hits column), see **Figure 28**. With TF2Network, we see that ABI5 (At2g36270) has experimentally determined protein-DNA interactions with 41% of the input set promoters. It is also coexpressed (based on TF2Network's own database of coexpression analyses) with 90% of the input gene set, and has PWM mappings to 16 input gene set promoters, making it a very good candidate as a potential regulator, as are a several other TFs at the top of the sorted Regulator Panel in **Figure 28**, like ABF4.



Regulator Panel displays the best potential TF with PWMs mapping to the input gene set promoters, and coexpression (CO) and PDI info (PD). Click the CO, PD or the PWM column head to sort on that column

Edge Manipulation Panel adjusts how and when the interactions are displayed

Cytoscape Panel shows the network visualization with genes as nodes and interactions as edges

Gene Info Panel allows for the export for interactions and various enrichments. Click "input_genes" to highlight the genes in the input set.

Regulator	Symbol	CO	PD	PWM	Rank	q-value	Hits
AT2G46270	GBF3	0%	49%	PWM951	25	2.52e-3	18
AT2G36270	ABI5	90%	41%	PWM2904 PWM621 PWM957	48 52 79	1.66e-2 1.53e-2 3.17e-2	16 16 15
AT4G01120	ATBZIP54	2%	37%	PWM1232	9	1.74e-3	19
AT4G34000	ABF3	0%	29%	PWM2529 PWM954	30 39	7.78e-3 5.99e-3	17 17
AT3G19290	ABF4	0%	25%	PWM966	6	1.27e-4	21
AT3G26790	FUS3	4%	10%	PWM917 PWM2165 PWM2726 PWM412 PWM518 PWM792	16 19 20 34 35 76	9.79e-4 3.32e-3 3.15e-3 6.87e-3 6.67e-3 3.29e-2	19 18 18 17 17 15
AT1G49720	ABF1	0%	4%	PWM974	59	1.35e-2	16
AT1G09530	PAP3	0%	2%	PWM829 PWM841	8 15	4.50e-4 1.04e-3	20 19
AT1G03970	GBF4	0%	0%	PWM994	26	2.42e-3	18

Export predictions

input_genes predicted_TFs all_genes

Gene ID	Symbol
AT1G03890	AT1G03890
AT1G04560	AT1G04560
AT1G05510	OBAP1a
AT1G14950	AT1G14950
AT1G17810	BETA-TIP
AT1G27461	AT1G27461
AT1G29680	AT1G29680
AT1G32560	AtLEA4-1

Export gene list

Gene Ontology | Protein-DNA | Protein-Protein | Co-expression

Biological Process		Molecular Function		Cellular Component	
lipid transport	2.	nutrient reservoir activity	2.	lipid particle	6.
seed development	3.	linoleic acid epoxygenase activity	4.	central vacuole	1.
oxylipin biosynthetic process	3.	plant seed peroxidase activity	4.	protein storage vacuole	2.
lipid storage	3.	heme oxygenase (decyclizing) activity	5.	extracellular region	3.
positive regulation of embryonic development	6.			primary cell wall	9.
				integral component of membrane	1.

Figure 28: Output of TF2Network for the 50 *ABI3*-coexpressed genes from Supplemental Table 1.

3.12 Targeting tools for confirming gene function

With many of the tools covered above, we might be able to predict a certain gene's involvement in some process we are interested in. Reverse genetics is undoubtedly one of the most powerful approaches to study gene function. This approach requires studying individuals with alterations in the gene of interest and seeking any phenotypes that may arise as a consequence of such alterations. Several strategies have been developed for this aim have been developed over the years, and Arabidopsis geneticists nowadays use mainly three tools for this purpose: (1) generating targeted mutations using the CRISPR-Cas9 system, (2) silencing of target genes using artificial microRNAs, or (3) taking advantage of the vast collection of T-DNA mutant lines.

3.12.1 CRISPR

Several strategies for genome editing currently exist, ranging from the use of zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), to the Clustered regularly interspaced short palindromic repeats- (CRISPR-) CRISPR associated protein 9 (Cas9) system. ZFNs were directly discarded as they remain difficult to engineer and prone to failure (Ramirez et al., 2008). TALENs were originally derived from the TAL effector proteins of tomato pathogen Xanthomonas, and reduced transformation efficiencies have raised concerns about potential TALEN cytotoxicity (Christian et al., 2013). In contrast, CRISPR-Cas9 has been repeatedly used in plants with virtually none of the presented drawbacks (Feng et al., 2013; Nekrasov et al., 2013).

To create different mutant alleles of a gene, we must first design guide RNAs (gRNAs). The gRNAs are aimed to recognize a DNA sequence in the genome containing a 3' protospacer adjacent motif (PAM), which would ultimately be targeted by the Cas9. Cleavage occurs around 3 bp upstream of the PAM domain, and those DNA breaks will eventually be repaired with deletions of nucleotides. In order to design spacer gRNA sequences specific to our gene of interest:

1. Go to https://www.genome.arizona.edu/crispr/CRISPRsearch.html. In the search box, search for "AT3G24650" and select "Arabidopsis thaliana" as the species.

2. The output will contain potential spacer sequences to be used for the gRNAs. These sequences are classified based on their potential off-target effects (Class0.0 being the lowest of all). The table also contains valuable information such as the exact genomic position of the targeting, which will be necessary for screening for mutations in plants exposed to these CRISPR-gRNAs.

3. Select a gRNA sequence by clicking on the empty box at the end of the row. This will expand the table and offer information about the cloning of this gRNA sequence, and the whether it is possible to use of restriction enzymes for cut analysis during genotyping. For cloning, a published plasmid vector (pRGE), containing necessary components to express Cas9 and gRNA, is suggested, and several other compatible alternatives exist.

3.12.2 Artificial miRNA

In certain cases, obtaining loss-of-function mutation for functional analysis of a gene is not a viable approach. Full knockouts might be lethal or we may need a more specific spatiotemporal reduction of the activity of our gene of interest. In such scenarios, the use of an artificial miRNA will be a valuable solution. The artificial microRNA (amiRNA) technology exploits endogenous miRNA precursors to generate sRNAs that direct gene silencing (Schwab et al., 2006). These amiRNAs are 21nt small RNAs that are genetically engineered to target and silence single or multiple genes (Ossowski et al., 2008). The level and spatiotemporal characteristics of the silencing ultimately depend on the expression profile of the amiRNA, allowing for more flexible frameworks for investigating gene function. The Web MicroRNA Designer 3 is an online app for the automated design of these artificial microRNAs.

1. Go to http://wmd3.weigelworld.org/. In the upper part, click on the “Designer” button. Make sure that the correct genome is selected (Araport11).

2. In the target gene box, write the Gene ID of your target gene (*e.g.* “*AT3G24650.1*”). Enter an email address to have your results sent to, and click “submit”. An email will be sent with a link to your result page.

3. The results show a summary of your job at the top, followed by a list of amiRNA sequences. The list contains the amiRNA sequence, the hybridization energy of the miRNA to a perfectly matching complement, the identifier of the target gene, and the hybridization energy of the miRNA to the target site in the target gene. These amiRNAs are ranked based on suitability and specificity, and the best ones in green.

4. Once selected, the amiRNA sequence needs to be engineered into an endogenous microRNA precursor for expression. The "Oligo" tool in the webpage helps design primers for modifying the *A. thaliana* MIR319a precursor. Click on the “Oligo” button, paste your selected miRNA sequence, select the RS300 vector, and click the “submit” button. This will provide the oligo sequences necessary to proceed with the cloning. Further information about the cloning strategy is located in the “Download” section of the webpage.

3.12.3 SIGnAL T-DNA Express

Perhaps one of the most powerful resources available to Arabidopsis geneticists are T-DNA mutant collections. These collections were created by large-scale *Agrobacterium tumefaciens* transfer-DNA (T-DNA)-induced insertional mutagenesis during the late 1990s and early 2000s (reviewed in O’Malley and Ecker, 2010). Because T-DNA inserts contain known sequences, primer design for genotyping assays is relatively straightforward and can allow for direct mapping to the genome to precisely identify the insertion locus. In the following steps, we will describe how to identify a SALK line mutant for *ABI3* and design its genotyping assay. While we primarily focus on the SALK collection, there are also the GABI-KAT, SAIL, and WISC lines available for T-DNA mutants. All of the mutant collections are in the Col-0 background. For a more detailed description of the T-DNA collections, please see O’Malley et al. (2015).

1. Go to http://signal.salk.edu/cgi-bin/tdnaexpress. You will see a genome browser showing gene models as green lines. Below the gene models, you will see graphical representations of genes and insertion mutants associated with that particular genomic region. Your goal here is to find T-DNA mutants for your gene of interest, *ABI3*. Under “1. Search”, enter “AT3G24650” as your query.

2. There is one SALK line of interest in the first exon of *ABI3,* SALK_138922. Generally speaking, when selecting insertional mutants, it is always best to select gene disruptions that are in the first exon, see **Figure 29**.

Legend denoting the T-DNA collections that are available. SALK lines are denoted with a green font.

***ABI3* locus,** along with other genes, are denoted in green at the top of the browser.

Position of T-DNA insertions are denoted in the lower portion of the browser. Here, we see that SALK_138922 has an insertion located in the first exon of *ABI3.*

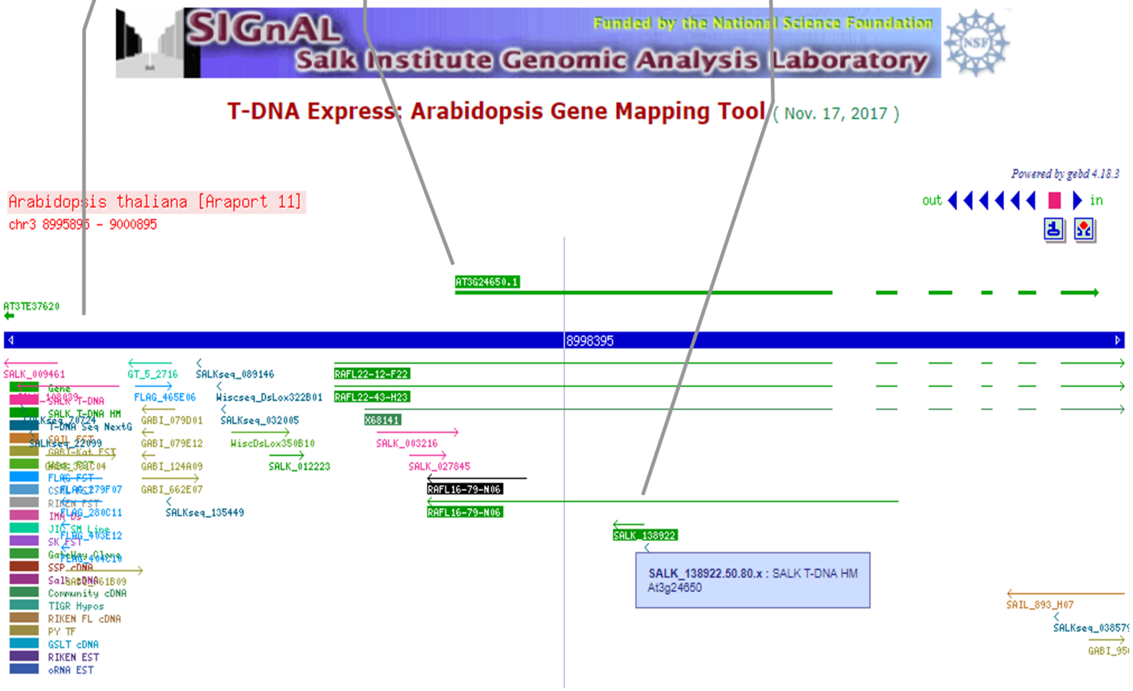


Figure 29: Screenshot of T-DNA Express genome browser at the *ABI3* locus. Genes and T-DNA insertions are displayed.

3. Click on "SALK_138922". You will be taken to a page with more detailed information about SALK_138922. In this case, the mutant is described as a knockout for *ABI3*, which is the most ideal case. To order this mutant, click on "Order from ABRC".

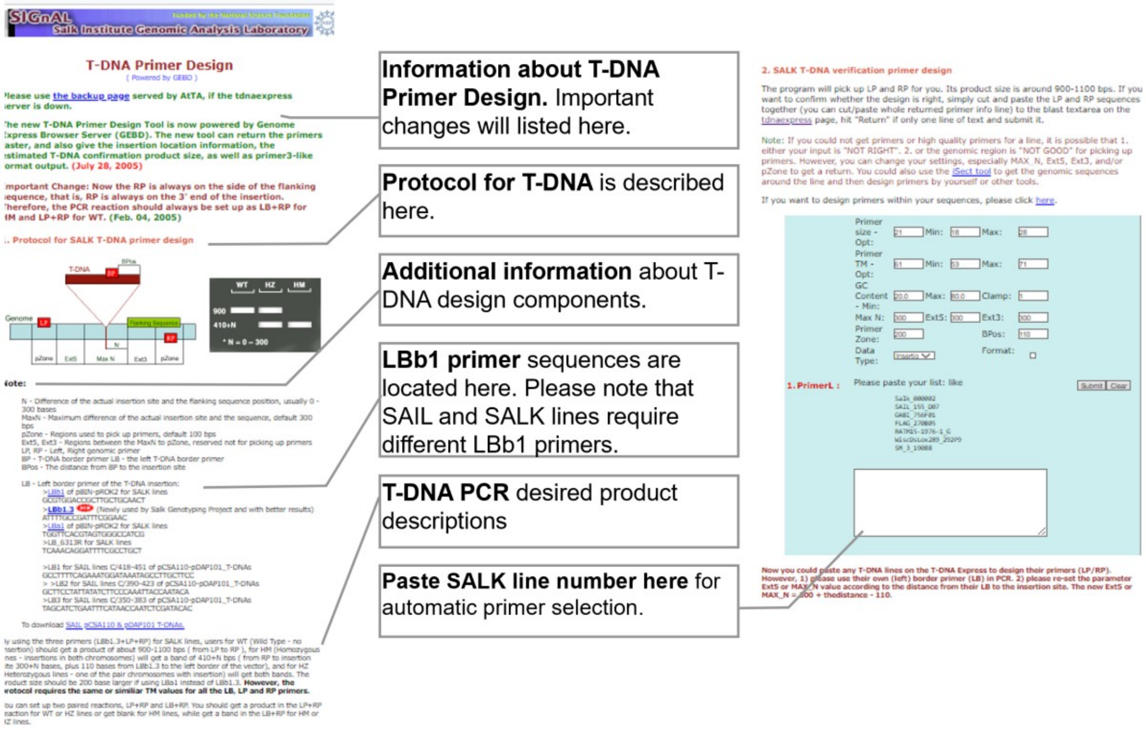
4. You will be taken to TAIR where additional information about this T-DNA line will be displayed. Some information worth noting about any mutant line are the name of parental line, ploidy number or whether the mutation is homozygous. To order, click on "Order from ABRC"[27].

[27] Note that the ABRC has recently rolled out its own interface for ordering T-DNA insertion lines and has disintermediated itself from TAIR, but the SIGnAL page still links to TAIR. Just click through as necessary.

Now that you have ordered your mutant, you should design genotyping primers for future experiments as well to verify that the mutant you ordered is the mutant you wanted.

1. Go to http://signal.salk.edu/tdnaprimers.2.html.

2. Here, the website displays the relatively straightforward goal of the genotyping assay you will design. You will order three primers, two of which you will design. Because the genotyping is based on insertion of a T-DNA in the region of interest, you can take advantage of this known sequence for genotyping. The primers that are for the endogenous (gene) sequences are denoted as LP and RP (left or right genomic primer). When used together in a wild-type background, these primers will produce a ~1-kb amplified fragment after PCR. When LP and RP are used in the mutant background, no amplified fragment should be produced (the T-DNA insertion is too long to amplify with a PCR extension time of one minute). To genotype mutant sequence of interest, you will use RP with a primer called LBb1 to amplify the right border of the T-DNA insertion. The primer LBb1 is designed specifically for the T-DNA insertion and can thus be used to genotype any SALK T-DNA line. When RP and LBb1 used to amplify mutant sequence, the resulting fragment should be about 410-bp, but could be as large 700-bp. Furthermore, if the primer melting temperatures are similar enough, all three primers can be used in a single genotyping PCR reaction, also known as a duplex reaction. Please note that for SAIL lines, a different LBb1 primer is used, as denoted below. In **Figure 30**, this information is displayed.



SIGnAL
Salk Institute Genomic Analysis Laboratory
Funded by the National Science Foundation

T-DNA Primer Design
(Powered by GEBD)

'lease use the backup page served by AtTA, if the tdnaexpress ierver is down.

'he new T-DNA Primer Design Tool is now powered by Genome :xpress Browser Server (GEBD). The new tool can return the primers aster, and also give the insertion location information, the :stimated T-DNA confirmation product size, as well as primer3-like ormat output. (July 28, 2005)

:mportant Change: Now the RP is always on the side of the flanking ;equence, that is, RP is always on the 3' end of the insertion. 'herefore, the PCR reaction should always be set up as LB+RP for IM and LP+RP for WT. (Feb. 04, 2005)

:. Protocol for SALK T-DNA primer design

Iote:

N - Difference of the actual insertion site and the flanking sequence position, usually 0 - 300 bases
MaxN - Maximum difference of the actual insertion site and the sequence, default 300 bps
pZone - Regions used to pick up primers, default 100 bps
Ext5, Ext3 - Regions between the MaxN to pZone, reserved not for picking up primers
LP, RP - Left, Right genomic primer
BP - T-DNA border primer LB - the left T-DNA border primer
BPos - The distance from BP to the insertion site

LB - Left border primer of the T-DNA insertion:
>LBb1 of pBIN-pROK2 for SALK lines
GCGTGGACCGCTTGCTGCAACT
>LBb1.3 (Newly used by Salk Genotyping Project and with better results)
ATTTTGCCGATTTCGGAAC
>LBa1 of pBIN-pROK2 for SALK lines
TGGTTCACGTAGTGGGCCATCG
>LB_6313R for SALK lines
TCAAACAGGATTTTCGCCTGCT

>LB1 for SAIL lines C/418-451 of pCSA110-pDAP101_T-DNAs
GCCTTTTCAGAAATGGATAAATAGCCTTGCTTCC
> >LB2 for SAIL lines C/390-423 of pCSA110-pDAP101_T-DNAs
GCTTCCTATTATATCTTCCCAAATTACCAATACA
>LB3 for SAIL lines C/350-383 of pCSA110-pDAP101_T-DNAs
TAGCATCTGAATTTCATAACCAATCTCGATACAC

To download SAIL pCSA110 & pDAP101 T-DNAs.

ly using the three primers (LBb1.3+LP+RP) for SALK lines, users for WT (Wild Type - no isertion) should get a product of about 900-1100 bps (from LP to RP), for HM (Homozygous nes - insertions in both chromosomes) will get a band of 410+N bps (from RP to insertion ite 300+N bases, plus 110 bases from LBb1.3 to the left border of the vector), and for HZ Heterozygous lines - one of the pair chromosomes with insertion) will get both bands. The roduct size should be 200 base larger if using LBa1 instead of LBb1.3. **However, the rotocol requires the same or similiar TM values for all the LB, LP and RP primers.**

ou can set up two paired reactions, LP+RP and LB+RP. You should get a product in the LP+RP eaction for WT or HZ lines or get blank for HM lines, while get a band in the LB+RP for HM or IZ lines.

Information about T-DNA Primer Design. Important changes will listed here.

Protocol for T-DNA is described here.

Additional information about T-DNA design components.

LBb1 primer sequences are located here. Please note that SAIL and SALK lines require different LBb1 primers.

T-DNA PCR desired product descriptions

Paste SALK line number here for automatic primer selection.

2. SALK T-DNA verification primer design

The program will pick up LP and RP for you. Its product size is around 900-1100 bps. If you want to confirm whether the design is right, simply cut and paste the LP and RP sequences together (you can cut/paste whole returned primer info line) to the blast textarea on the tdnaexpress page, hit "Return" if only one line of text and submit it.

Note: If you could not get primers or high quality primers for a line, it is possible that 1. either your input is "NOT RIGHT". 2. or the genomic region is "NOT GOOD" for picking up primers. However, you can change your settings, especially MAX_N, Ext5, Ext3, and/or pZone to get a return. You could also use the iSect tool to get the genomic sequences around the line and then design primers by yourself or other tools.

If you want to design primers within your sequences, please click here.

Primer size - Opt: 21 Min: 18 Max: 28
Primer TM - Opt: 61 Min: 53 Max: 71
GC Content - Min: 20.0 Max: 80.0 Clamp: 1
Max N: 300 Ext5: 300 Ext3: 300
Primer Zone: 200 BPos: 110
Data Type: Insertio Format: □

1. PrimerL : Please paste your list: like
Salk_000002
SAIL_155_D07
GABI_756F01
FLAG_270B05
RATM15-1976-1_G
WiscDsLox289_292P9
SM_3_19088

Now you could paste any T-DNA lines on the T-DNA Express to design their primers (LP/RP). However, 1) please use their own (left) border primer (LB) in PCR. 2) please re-set the parameter Ext5 or MAX_N value according to the distance from their LB to the insertion site. The new Ext5 or MAX_N = 300 + thedistance - 110.

Figure 30: T-DNA primer design for mutant genotyping.

3. Go to “2. SALK T-DNA verification primer design”. To design your primers of interest, select primer size (*e.g.* 20-bp), primer melting temperature (Tm) (*e.g*. 61°C), and GC content (*e.g*. 20%). Additional parameters of importance are the Max N (maximum difference of the actual insertion site and the sequence), as well as Ext5 and Ext3 (regions between the MaxN to pZone, reserved for not picking primers). Here, it is best to keep parameters at the defaults.

4. Next, paste in the number for your SALK line interest: “SALK_138922”. Click “submit”.

5. You will then be taken to a page displaying the following information:
   ```
   SALK_138922.50.80.x  PRODUCT_SIZE 1253  PAIR_ANY_COMPL 0.00
   PAIR_3'_COMPL 0.00  DIFF_TM 0.21  LP TCGGTCCATGGTAGGTAACTG  Len 21
   TM 59.86  GC 52.38  SELF_ANY_COMPL 0.21  3'_COMPL 0.00 RP
   GAGAAGATCCGACTCCAAACC  Len 21  TM 60.07  GC 52.38  SELF_ANY_COMPL
   0.21  3'_COMPL 0.00  Insertion chr3 8998753  BP+RP_PRODUCT_SIZE 579-879
   ```
   As you can see, all of the relevant information is available for ordering genotyping primers, such as sequence and melting temperatures. Here, the Tm’s are similar enough for all three primers being used in a single duplex PCR reaction.

4 Results/Discussion

In summary, it is clear that there is an enormous volume of data at the disposal of *Arabidopsis* researchers, and by accessing it with the tools described in this lab and others, insights can be garnered that will help with elucidating the role of a gene or set of genes in question. As a final note, the International Arabidopsis Informatics Consortium (IAIC, 2012) has worked with developers and researchers to launch an Arabidopsis Information Portal, Araport.org, which brings together the data described for the tools above and from elsewhere into a seamless interface. Data is constantly being added to this database and to many of the databases described in this chapter, so it is worthwhile to check them all frequently.

5 References

1001 Genomes Consortium (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell **166**: 481–491.

Alonso, J.M. et al. (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science **301**: 653–657.

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. (2016). ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression. Plant Cell Physiol. **57**: e5–e5.

Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. **25**: 25–29.

Austin, R.S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T.T., Fan, J., Foong, C., Breit, R., Desveaux, D., Moses, A., and Provart, N.J. (2016). New BAR tools for mining expression data and exploring Cis-elements in Arabidopsis thaliana. Plant J. Cell Mol. Biol.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. **37**: W202-208.

Baud, S., Boutin, J.-P., Miquel, M., Lepiniec, L., and Rochat, C. (2002). An integrated overview of seed development in Arabidopsis thaliana ecotype WS. Plant Physiol. Biochem. **40**: 151–160.

Brady, S.M. et al. (2011). A stele-enriched gene regulatory network in the Arabidopsis root. Mol. Syst. Biol. **7**: 459.

Brady, S.M., Orlando, D.A., Lee, J.-Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A High-Resolution Root Spatiotemporal Map Reveals Dominant Expression Patterns. Science **318**: 801–806.

Brady, S.M. and Provart, N.J. (2009). Web-Queryable Large-Scale Data Sets for Hypothesis Generation in Plant Biology. Plant Cell **21**: 1034–1051.

Brady, S.M., Sarkar, S.F., Bonetta, D., and McCourt, P. (2003). The ABSCISIC ACID INSENSITIVE 3 (ABI3) gene is modulated by farnesylation and is involved in auxin signaling and lateral root development in Arabidopsis. Plant J. **34**: 67–75.

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

Chatr-Aryamontri, A. et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. **45**: D369–D379.

Chory, J. 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 Physiol. **123**: 423–426.

Christian, M., Qi, Y., Zhang, Y., and Voytas, D.F. (2013). Targeted Mutagenesis of Arabidopsis thaliana Using Engineered TAL Effector Nucleases. G3 Genes Genomes Genet. **3**: 1697–1705.

Cui, H., Levesque, M.P., Vernoux, T., Jung, J.W., Paquette, A.J., Gallagher, K.L., Wang, J.Y., Blilou, I., Scheres, B., and Benfey, P.N. (2007). An Evolutionarily Conserved Mechanism Delimiting SHR Movement Defines a Single Layer of Endodermis in Plants. Science **316**: 421–425.

Dill, A., Jung, H.S., and Sun, T.P. (2001). The DELLA motif is essential for gibberellin-induced degradation of RGA. Proc. Natl. Acad. Sci. U. S. A. **98**: 14162–14167.

Dong, S. et al. (2019). Proteome-wide, Structure-Based Prediction of Protein-Protein Interactions/New Molecular Interactions Viewer. Plant Physiol. **179**: 1893–1907.

Dubreucq, B., Berger, N., Vincent, E., Boisson, M., Pelletier, G., Caboche, M., and Lepiniec, L. (2000). The Arabidopsis AtEPR1 extensin-like gene is specifically expressed in endosperm during seed germination. Plant J. Cell Mol. Biol. **23**: 643–652.

Feng, Z., Zhang, B., Ding, W., Liu, X., Yang, D.-L., Wei, P., Cao, F., Zhu, S., Zhang, F., Mao, Y., and Zhu, J.-K. (2013). Efficient genome editing in plants using a CRISPR/Cas system. Cell Res. **23**: 1229–1232.

Finkelstein, R.R. and Somerville, C.R. (1990). Three Classes of Abscisic Acid (ABA)-Insensitive Mutations of Arabidopsis Define Genes that Control Overlapping Subsets of ABA Responses. Plant Physiol. **94**: 1172–1179.

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

Gaudinier, A. et al. (2011). Enhanced Y1H assays for Arabidopsis. Nat. Methods **8**: 1053–1055.

Geisler-Lee, J., O’Toole, N., Ammar, R., Provart, N.J., Millar, A.H., and Geisler, M. (2007). A predicted interactome for Arabidopsis. Plant Physiol. **145**: 317–329.

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

Higo, K., Ugawa, Y., Iwamoto, M., and Higo, H. (1998). PLACE: a database of plant cis-acting regulatory DNA elements. Nucleic Acids Res. **26**: 358–359.

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

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv. Bioinforma. **2008**: 420747.

Hubbard, T. et al. (2002). The Ensembl genome database project. Nucleic Acids Res. **30**: 38–41.

IAIC (2019). Arabidopsis bioinformatics resources: The current state, challenges, and priorities for the future. Plant Direct **3**: e00109.

IAIC (2012). Taking the Next Step: Building an Arabidopsis Information Portal. Plant Cell Online **24**: 2248–2256.

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

Katari, M.S., Nowicki, S.D., Aceituno, F.F., Nero, D., Kelfer, J., Thompson, L.P., Cabello, J.M., Davidson, R.S., Goldberg, A.P., Shasha, D.E., Coruzzi, G.M., and Gutiérrez, R.A. (2010). VirtualPlant: A Software Platform to Support Systems Biology Research. Plant Physiol. **152**: 500–515.

Kersey, P.J. et al. (2018). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. **46**: D802–D808.

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

Krishnakumar, V. et al. (2015). Araport: the Arabidopsis Information Portal. Nucleic Acids Res. **43**: D1003–D1009.

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

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

Li, B., Gaudinier, A., Tang, M., Taylor-Teeples, M., Nham, N.T., Ghaffari, C., Benson, D.S., Steinmann, M., Gray, J.A., Brady, S.M., and Kliebenstein, D.J. (2014). Promoter-based integration in plant defense regulation. Plant Physiol. **166**: 1803–1820.

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

de Lucas, M., Davière, J.-M., Rodríguez-Falcón, M., Pontin, M., Iglesias-Pedraz, J.M., Lorrain, S., Fankhauser, C., Blázquez, M.A., Titarenko, E., and Prat, S. (2008). A molecular framework for light and gibberellin control of cell elongation. Nature **451**: 480–484.

de Lucas, M., Pu, L., Turco, G., Gaudinier, A., Morao, A.K., Harashima, H., Kim, D., Ron, M., Sugimoto, K., Roudier, F., and Brady, S.M. (2016). Transcriptional Regulation of Arabidopsis Polycomb Repressive Complex 2 Coordinates Cell-Type Proliferation and Differentiation. Plant Cell **28**: 2616–2631.

McCarty, D.R., Carson, C.B., Stinard, P.S., and Robertson, D.S. (1989). Molecular Analysis of viviparous-1: An Abscisic Acid-Insensitive Mutant of Maize. Plant Cell **1**: 523–532.

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

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

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

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

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

Mu, J., Tan, H., Zheng, Q., Fu, F., Liang, Y., Zhang, J., Yang, X., Wang, T., Chong, K., Wang, X.-J., and Zuo, J. (2008). LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis. Plant Physiol. **148**: 1042–1054.

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

Murphy, E. et al. (2016). RALFL34 regulates formative cell divisions in Arabidopsis pericycle during lateral root initiation. J. Exp. Bot. **67**: 4863–4875.

Nakabayashi, K., Okamoto, M., Koshiba, T., Kamiya, Y., and Nambara, E. (2005). Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. Plant J. Cell Mol. Biol. **41**: 697–709.

Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J.D.G., and Kamoun, S. (2013). Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. Nat. Biotechnol. **31**: 691–693.

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

Nole-Wilson, S., Tranby, T.L., and Krizek, B.A. (2005). AINTEGUMENTA-like (AIL) genes are expressed in young tissues and may specify meristematic or division-competent states. Plant Mol. Biol. **57**: 613–628.

Obayashi, T. and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes **16**: 249–260.

O'Malley, R.C., Barragan, C.C., and Ecker, J.R. (2015). A user's guide to the Arabidopsis T-DNA insertion mutant collections. Methods Mol. Biol. Clifton NJ **1284**: 323–342.

O'Malley, R.C. and Ecker, J.R. (2010). Linking genotype to phenotype using the Arabidopsis unimutant collection. Plant J. Cell Mol. Biol. **61**: 928–940.

O'Malley, R.C., Huang, S.-S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292.

Orchard, S. et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. **42**: D358-363.

Ossowski, S., Schwab, R., and Weigel, D. (2008). Gene silencing in plants using artificial microRNAs and other small RNAs. Plant J. Cell Mol. Biol. **53**: 674–690.

Porco, S. et al. (2016). Lateral root emergence in Arabidopsis is dependent on transcription factor LBD29 regulation of auxin influx carrier LAX3. Dev. Camb. Engl. **143**: 3340–3349.

Provart, N. and Zhu, T. (2003). A browser-based functional classification SuperViewer for Arabidopsis genomics. Curr. Comput. Mol. Biol. **2003**: 271–272.

Ramirez, C.L. et al. (2008). Unexpected failure rates for modular assembly of engineered zinc fingers. Nat. Methods **5**: 374–375.

Reiser, L., Subramaniam, S., Li, D., and Huala, E. (2017). Using the Arabidopsis Information Resource (TAIR) to Find Information About Arabidopsis Genes. Curr. Protoc. Bioinforma. **60**: 1.11.1-1.11.45.

Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-Throughput Sequencing Technologies. Mol. Cell **58**: 586–597.

Roudier, F. et al. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. **30**: 1928–1938.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of Arabidopsis thaliana development. Nat. Genet. **37**: 501–506.

Schwab, R., Ossowski, S., Riester, M., Warthmann, N., and Weigel, D. (2006). Highly Specific Gene Silencing by Artificial MicroRNAs in Arabidopsis. Plant Cell **18**: 1121–1133.

Sparks, E.E. et al. (2016). Establishment of Expression in the SHORTROOT-SCARECROW Transcriptional Cascade through Opposing Activities of Both Activators and Repressors. Dev. Cell **39**: 585–596.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell **152**: 352–364.

Sullivan, A.M., Bubb, K.L., Sandstrom, R., Stamatoyannopoulos, J.A., and Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. Curr. Plant Biol. **3–4**: 40–47.

Taylor-Teeples, M. et al. (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517**: 571–575.

The EPIC Planning Committee, T.E.P. (2012). Reading the Second Code: Mapping Epigenomes to Understand Plant Growth, Development, and Adaptation to the Environment. Plant Cell **24**: 2257–2261.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. Cell Mol. Biol. **37**: 914–939.

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

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

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., and Provart, N.J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. **32**: 1633–1651.

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

Waese, J. et al. (2017). ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology. Plant Cell **29**: 1806–1821.

Warde-Farley, D. et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. **38**: W214-220.

Willems, P., Horne, A., Parys, T.V., Goormachtig, S., Smet, I.D., Botzki, A., Breusegem, F.V., and Gevaert, K. (2019). The Plant PTM Viewer, a central resource for exploring plant protein modifications. Plant J. **[epub ahead of print]**.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and 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., and Yang, Y. (2014). Genome-Wide Prediction of Highly Specific Guide RNA Spacers for CRISPR–Cas9-Mediated Genome Editing in Model Plants and Major Crops. Mol. Plant **7**: 923–926.

Yang, W., Zhang, W., and Wang, X. (2017). Post-translational control of ABA signalling: the roles of protein phosphorylation and ubiquitination. Plant Biotechnol. J. **15**: 4–14.

Yao, Q., Bollinger, C., Gao, J., Xu, D., and Thelen, J.J. (2012). P(3)DB: An Integrated Database for Plant Protein Phosphorylation. Front. Plant Sci. **3**: 206.

Yao, Q., Ge, H., Wu, S., Zhang, N., Chen, W., Xu, C., Gao, J., Thelen, J.J., and Xu, D. (2014). P^3DB 3.0: From plant phosphorylation sites to protein networks. Nucleic Acids Res. **42**: D1206-1213.