Leveraging coexpression improves cross-species comparison and enables single-cell integration between distant plant species

MICHAEL JOHN PASSALACQUA

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

School of Biological Sciences

Cold Spring Harbor Laboratory

September 11, 2024

"The happiness of your life depends upon the quality of your thoughts."

Marcus Aurelius

Acknowledgements

First, I want to thank my advisor, Dr. Jesse Gillis, for his fantastic mentorship during my Ph.D. He has been nothing but patient and nurturing, and his deep care for all his trainees shines through in the dedication, support, and time he commits to all of them. Although I had no computational experience at the start of my Ph.D., he was willing to take me into his lab and allow me to work on plant research, providing me with the environment I needed to grow and flourish.

I would also like to thank my committee members, Dr. Peter Koo, Dr. Thomas Gingeras, and Dr. Zachary Lippman (also Director of CSHL Ph.D. program, for which he receives additional thanks). Their insights and advice during committee meetings have been critical to shaping and improving my work. Additionally, I would like to thank everyone at the CSHL School of Biological Sciences – Alsyon Kass-Eisler, Kimberly Creteur, Catherine Perez, Monn Monn Myat, and the emeritus dean, Alex Gann. Their support and patience with my late forms has been critical.

The Gillis Lab members, now spread across the world, deserve special thanks. Although we have been remote for several years now, their support and friendship have eased my journey substantially. Additionally, I would like to thank the Meyer and Koo labs for adopting me socially following the departure of the Gillis Lab from CSHL. And of course, thanks to all my friends and family whose support and friendship have kept me sane.

Contents

List of Figures

Dedicated to my fountain of joy, Brooke

Chapter 1

Introduction to genetic diversity in plants and coexpression

Plants are a highly diverse set of organisms upon which almost all human food production depends. Understanding how genes and gene modules in diverse flowering plants (angiosperms) relate to each other and improving these comparisons is critical to being able to understand the mechanisms underlying the desirable traits of different plant species – the first step towards transferring or mimicking these traits in crop species of interest. However, the rapid expansion of plant genomes complicates cross-species comparisons, making attempts to compare plants and transfer knowledge from traditional model systems to crop species difficult. This chapter establishes the background necessary to understand how we can improve cross-species comparisons in plants using coexpression. We begin by introducing plant science and the importance of being able to make cross-species comparisons to the food supply (Chapter 1.1), then introduce the impact of whole genome duplications (WGD) (Chapter 1.2). Finally, we discuss the history of coexpression analysis (Chapter 1.3) and innovations in network analysis that leverage coexpression networks (Chapter 1.4), and why they are especially useful in plants.

1.1 The expansive diversity of plants

In 1879 Darwin wrote, discussing the fossil record of plants, that "the rapid development as far as we can judge of all the higher plants within recent geological times is an abominable mystery" (Darwin 1903). The diversification of angiosperms was unusually large and fast, and the sparse macrofossil record left as a result was a source of confusion, given Darwin's familiarity with a more gradual appearance of species. Progress towards unraveling this abominable mystery was slow, until advances in palynology (the study of fossilized pollen) allowed the differentiation of species without leaf records (Doyle 1978; Hickey and Doyle 1977). Today, there are almost 300,000 species of angiosperms, vastly outnumbering the gymnosperms, of which there are only 1000(Christenhusz and Byng 2016). The species resulting from this radiation event were tremendously successful, dominating most habitats (Crepet and Niklas 2009). The rapid colonization of so many unique habitats resulted in an extremely diverse set of organisms — ranging from towering 300 foot *Eucalyptus* trees on the island of Tasmania to microscopic *Wolffia* common in local ponds (Sree, Sudakaran, and Appenroth 2015; Waters, Burrows, and Harper 2010). In one extreme example, the silverswords diverged into small flat plants, vines, and trees in only 6 million years (Witter and Carr 1988). This diversity is not limited to morphology, as angiosperms are also highly diverse in stress tolerance (with examples of species highly tolerant of aridity, extreme temperatures, or high salinity) and biochemical diversity (Folk, Siniscalchi, and

Soltis 2020). Some groups, such as *Brassicacea*, produce novel secondary metabolites like glucosinolates to deter pests and predators, while others, such as *Orchidaceae* produce terpenes to attract insect pollinators (Teoh 2015). The broad diversity of angiosperms was critical to the development of human agriculture, and they now make up the vast majority of human crops. All cereal crops, almost all culinary fruits, most vegetables, and most nuts are angiosperms. Today, farmland covers 4.62 billion acres of land, and helps feed over 8 billion people (Oliphant, Thenkabail, and Teluguntla 2022). But in 1798, Malthus predicted that as the human population growth would inevitably outstrip the ability of the land to support this population, claiming that "The food therefore which before supported seven millions, must now be divided among seven millions and a half or eight millions. The poor consequently must live much worse, and many of them be reduced to severe distress" (Malthus 1809). Clearly, he was wrong. Later biologists, like Paul R. Ehrlich also predicted mass starvation in India and across the globe (Ehrlich 1983). Yet, neither Malthus nor Paul R. Elrich anticipated the massive impact the Green Revolution would have on agricultural production. In Asia, the total yield from 1961 to 1980 increased by over 3-fold (Evenson and Gollin 2003). As a result of work by Norman Borlaug and countless other plant scientists and breeders, over a billion people were spared starvation, thanks primarily to the introduction of modern, high yield varieties of dwarf wheat and rice.

In the decades since the Green Revolution, plant research has only accelerated, and the development of molecular breeding techniques greatly

increased the speed of plant breeding. One of the first modern developments was that of molecular marker systems, which used the linkages of desirable traits with different isozymes to map and then select key traits of interest (Edwards, Stuber, and Wendel 1987; Brewer, Sing, and Sears 1969). These techniques, invented prior to availability of whole genome sequencing, associated regions of a chromosome with loci that improved performance. Being able to confirm the presence of loci in a cross improved the ability of scientists to rapidly introgress traits into existing elite lines. While prior maps used morphological variation, these morphological traits were often deleterious to yield. By instead using the tiny differences between isozymes, breeders were able to map with a non-deleterious marker (*Isozymes in Plant Genetics and Breeding* 2012). Next, the development of transgenic crops powered by *Agrobacterium* transformation enabled the addition of traits not present in existing plants. In 1983, multiple labs published work describing the creation of transgenic plants via *Agrobacterium* transformation (Fraley et al. 1983; Hoekema et al. 1983; Herrera-Estrella et al. 1983; Bevan, Flavell, and Chilton 1983). The unprecedented ability to add any gene to a complex, multicellular organism opened the door to completely new avenues of plant breeding. The first approved transgenic crop, the FlavrSavr tomato, contained two added genes (Fernandez-Cornejo et al. 2014). The selection gene encoded resistance to antibiotics, and the second gene, an antisense gene, inhibited the expression of a native enzyme that broke down pectin, softening the tomato. While the FlavrSavr was not particularly successful, other transgenic crops, especially those encoding BT toxins for resistance to pests or tolerance for the herbicide glyphosate, were

wildly successful. Today, nearly all corn, cotton, and soybeans grown in the United States are transgenic (Dodson 2024).

Although the Green Revolution and the research that followed was wildly successful, the benefits of this research have not been as equally felt. The vast majority of research occurs in a handful of tractable model organisms, and yield benefits are often limited to these organisms and their close relatives. The biggest winners have been cereals like wheat, maize, and rice which have experienced unprecedented gains in yield over the last 100 years (Evenson and Gollin 2003). In comparison to regions like Asia, areas such as the Middle East and Africa have experienced much more modest gains in yield (Evenson and Gollin 2003). While this situation has many causes, including underinvestment in mechanized agriculture, one contribution is the lack of investment into research of crops grown in African nations. While this situation is slowly shifting, much work remains to be done to transfer the decades of accumulated knowledge in model organisms like Arabidopsis, tomato, and maize to the panoply of other crops grown across the world.

Transferring knowledge from model species to distantly related relatives is a critical goal in plant science, but major barriers have prevented the smooth transition of knowledge between species. Until relatively recently, sequencing a new genome was a major undertaking that required substantial funding (Wetterstrand 2023). As sequencing costs have come down, individual labs have increasingly become able to sequence and annotate their own genomes (Armstrong

et al. 2019). The increase in accessibility of genomics brought on by this shift has resulted in a flood of new model organisms for studying unique and valuable traits and enabling cross-species comparisons of morphological, developmental, and biochemical traits (Zhuang and Zhang 2021; Hilgenhof et al. 2023; Messeder et al. 2024; Schuster et al. 2024). Despite this surge in resources, cross-species comparisons in plants remain more difficult than in other clades, such as Animalia (Plant Cell Atlas Consortium et al. 2021).

Several factors contribute to the difficulty in making cross-species comparisons between plant species. The first is that the rapid diversification of plant morphology often muddies tissue and developmental homology between species. When making a comparison between two closely related species, like kohlrabi and turnip, it may not be clear what tissue in kohlrabi corresponds to the turnip tuber, the main organ of interest (Hearn, O'Brien, and Poulsen 2018). Should you compare the stem, which is the storage tissue for kohlrabi, or the root? For more distantly related species, this issue may be even more severe. Additionally, as species diverge their life plans timing may also diverge. It is unclear how you would correctly stage tissues when comparing an annual species to one that is biannual or perennial. Even for species with similar lifespans, different parts of development may occur at different rates in each species (Calderwood et al. 2021). This uncoupling makes it difficult to properly stage tissues, as correctly staging for one aspect of development may desync another. Secondly, plant genomes tend to evolve more rapidly (Clark and Donoghue 2018). By comparing chromosome structure between grasses (Poaceae) (60 million years

of evolution) and primates (90 million years of evolution), one can clearly identify the accelerated change in chromosome structure in plants (Zhao and Schranz 2019). While primates (and mammals as a whole) are highly syntenic, Poaceae (and angiosperms as a whole) have very little conserved synteny, indicating massive rearrangements of the genome. This is likely a result of plant tolerance to massive genomic shocks, like whole genome duplications and chromosomal fusions (Panchy, Lehti-Shiu, and Shiu 2016). This tolerance also influences our third factor, that plant gene families tend to expand much more rapidly than animal gene families (Rensing et al. 2008). There are three main sources for these gene family expansions: tandem arrays, transposons, and whole genome duplications (Flagel and Wendel 2009). Tandem arrays are sections of the genome where the same gene is duplicated multiple times, typically because of unequal chromosomal crossover during recombination, and may be up to 15% of the genome (Jander and Barth 2007). Transposons, which are more active in plants than mammals, may accidentally bring along additional genes with them when copying themselves (Lisch 2013). Finally, whole genome duplications typically occur from errors during meiosis, where a gamete fails to reduce its copy number (Clark and Donoghue 2018). We will pay particular attention to whole genome duplications, as they are uniquely frequent in plants, and have key properties that influence plant evolution.

1.2 Whole genome duplications and their impact on plant diversification

Today, we know that critical whole genome duplication events co-occurred with the angiosperm radiation event (Clark and Donoghue 2018). Although it is difficult to assign causality to these duplication events, as other plant lineages have experienced whole genome duplication and not undergone dramatic radiation, it seems likely that the whole genome duplication events were an important factor to the rapid radiation of angiosperms. The potential mechanisms by which a whole genome duplication can enable radiation are often synergistic, enabling rapid evolutionary shifts. Briefly, I will outline these mechanisms and their importance to plant evolution.

One of the primary contributions of a whole genome duplication is that it doubles the amount of genetic material, loosening selection constraints on paralog pairs. As described by Ohno, having two copies of a gene enables one copy to retain the ancestral function that may be critical to the organism fitness, freeing the second copy to diverge and pick up a different function (Holland 1999). These changes can typically be categorized into two different types of gene modifications. During subfunctionalization, a mutation in the gene or its regulatory region causes the gene to be capable of only a subset of what it was previously capable of (Birchler and Yang 2022). Global expression might shift to tissue specific expression, or a domain might be rendered non-functional, causing a normally catalytic protein to instead sequester its target. The less common change

in genes is neofunctionalization, when a copy of the original gene develops a new function (Birchler and Yang 2022). Because mutations are more likely to disrupt function than to enhance it, neofunctionalizations are less common than sub-functionalization. Occasionally, a sub-functionalized gene will later develop a new function, a process termed subneofunctionalization. Additionally, some fates, such as hypofunctionalization and compensatory drift, do not result in new functions for either gene (Birchler and Yang 2022).

Second, copying the entire genome enables entire gene modules to diverge together. Instead of having a single gene freed from selection constraint, an entire gene module is freed, providing the opportunity for the coexpressed genes to evolve a new function together. This is valuable, because while duplications on single genes can generate the genetic redundancy necessary for innovation, genes often act in concert with other genes to collectively accomplish a goal. In these cases, the genes are often regulated together, forming a regulatory module of genes that express at the same time. For these sets of genes, a single gene duplication does not provide the necessary redundancy for regulatory rewiring (Clark and Donoghue 2018). Instead, a whole genome duplication is necessary to provide the genetic space for innovations on the existing regulatory wiring. These genes often already have shared regulatory machinery, such as shared transcription factor binding sites and motifs, further priming them for co-evolution.

Third, work by James Birchler and others has shown that duplications of all chromosomes are much more tolerated than duplications of individual

chromosomes (Satina, Blakeslee, and Avery 1937; Blakeslee 1934; Birchler 2013). Genes on the same chromosome are often coexpressed, so one might anticipate an aneuploidy having a similar effect as a whole genome duplication. However, these events often fail to produce viable, diverse progeny. Besides the impact of aneuploidy on successful meiosis, these progenies tend to be sickly, while progeny of a whole genome duplication are healthy. This is because in whole genome duplications, the dosage balance between all genes are maintained. Because the amount of all mRNA produced is doubled, all proteins are produced in approximately the same ratio, preventing deleterious effects from unbalanced interactions between proteins (Edger and Pires 2009). In contrast, only some genes are duplicated in an aneuploidy, so if they need to interact with another gene that was not duplicated, there will be half as much of it, relatively. Confirming this, aneuploidy in species with higher base ploidy levels has less of an impact, because gene balance is disturbed to a lesser degree (Sears 1953). In contrast, following a whole genome duplication, gene dosage balance is maintained — the ratio of the amount of each gene product is the same. Because of this, polyploid plants are often indistinguishable in phenotype from their parents, or they may simply be slightly larger (Corneillie et al. 2019). Aneuploid plants are clearly distinguishable from their parents, and often respond more poorly to stress. This contrasts with mammals, where whole genome duplications are typically lethal, and as such whole genome duplications are quite rare. One potential explanation for this difference is that plant development is stochastic, with no set body plan, while animal development is tightly regulated (Kejnovsky, Leitch, and Leitch 2009). The

plasticity of plants is also relevant to their germ line, which is not sequestered, as it is in mammals. As such, the plant germ line must be much more tolerant to mutations (Fajkus, Sýkorová, and Leitch 2005).

One of the only recent whole genome duplications occurred in salmonids, and has been investigated to better understand whole genome duplications in animals (Berthelot et al. 2014). This duplication occurred 88 million years ago (MYA), and did not immediately precede a radiation in salmonids, which occurred 45 million years later (Macqueen and Johnston 2014). Other recent whole genome duplications in animals include recent events in carp and Xenopus (Z. Chen et al. 2019; Session et al. 2016). The tolerance of plants to whole genome duplications, but not to aneuploidy, means that whole genome duplications uniquely position plants to undergo the previously mentioned co-divergence of entire gene modules.

Finally, the polyploidization associated with whole genome duplications throws up instant barriers to gene flow, creating an "instant speciation" where the polyploid is forced to rely on selfing to produce a new generation (Clark and Donoghue 2018). The two main types of polyploid are autopolyploids, which form when a single species doubles its own genome, and allopolyploids, which often arise via hybridization of two different species combined with a genome doubling. In both cases, the diploid (or higher ploidy) gametes are usually incompatible with the haploid gametes of the parents. This incompatibility prevents successful fertilization, effectively terminating gene flow between the polyploid and its progenitors. This reproductive isolation is a key factor in the speciation process, as

it allows the new polyploid lineage to evolve independently. However, as many plants are capable of selfing, the new polyploid is still able to spread, giving time for the population to grow and eventually undergo rediploidization. Both types of polyploids may also experience masking of deleterious recessive alleles by the additional gene copies (Ren et al. 2018). Uniquely in allopolyploids, the two sets of chromosomes will typically be maintained as separate sub-genomes inherited by all progeny. Because of this, the heterosis experienced by the progeny is inheritable, providing a permanent boost in fitness (Bansal, Banga, and Banga 2012; Madlung 2013). This effect may contribute to the frequency of allopolyploid crops.

The unique robustness of plants to genetic perturbation due to gene duplication, as well as their rapid diversification, makes plants an excellent model in which to investigate the divergence, rewiring, and repurposing of genes between species. Understanding how genes specify cell types, and perform new functions requires cross-species comparisons. However, the same traits that make plants a good model system to study gene rewiring also introduces substantial challenges to capturing shifts in gene function. Frequent gene family expansion from whole genome duplications and other local expansions as well as unclear homology between morphologically distinct species makes simple, expression-based comparisons difficult. It is often unclear with what two genes you should compare, and in what tissues and at what time. Additionally, while some functional shifts in genes are driven by changes in their sequence and are easily captured by traditional sequencing, other shifts in function occur when changes in regulatory regions

affect a gene's expression. These regulatory shifts cause the gene to diverge from a paralog without a change in the sequence of the gene, changing its function (e.g. during subfunctionalization). To instead capture shifts in both gene biochemical function and regulation, we turn to coexpression, which associates genes based on how tightly their expression is linked.

1.3 Constructing and comparing coexpression networks

The concurrent development of microarray technology and release of the first complete genome sequence of a eukaryote (*Saccharomyces cerevisiae)* in the mid-1990s enabled the first high throughput measurements of mRNA levels in bulk samples (Lashkari et al. 1997; Schena et al. 1995). Prior to this, the main way to capture and quantify the amount of mRNA present in a sample was a Northern blot. Northern blots require first running an mRNA sample out onto a gel before transferring it to a membrane and visualizing it with a probe hybridized to a detection system (Alwine, Kemp, and Stark 1977). Each RNA visualized requires a probe and overlapping RNAs on the gel (which separates them by size) cannot be visualized simultaneously. In contrast, microarrays can measure the expression of thousands of genes at once. To accomplish this, tiny dots containing probes for each gene are deposited onto a glass slide (Heller 2002). Then, during preparation of the sample libraries, the cDNA/RNA is labeled with a fluorescence system, to quantify the amount of mRNA present via the fluorescence of each dot. The fluorescence of each dot is higher as the amount of mRNA increases. The ability

to measure the expression of thousands of genes in parallel also enabled the calculation of coexpression scores between them, generating coexpression networks for much or all of the genome. As sequencing technology became more advanced, microarrays were gradually phased out in favor of bulk RNA-sequencing (RNA-seq) technology, which provides direct counts of the mRNA molecules present, instead of a relative abundance measurement (Stark, Grzelak, and Hadfield 2019). Further advancements in chemistry and microfluidics eventually enabled the sequencing of mRNA from individual cells, an advancement that unlocked the creation of more specialized and specific coexpression networks (Tang et al. 2009; Olsen and Baryawno 2018). I will highlight key papers in the field that constructed and compared coexpression networks using at the time state of the art methods for generating expression data.

Utilizing some of the first microarray technology, Eisen published a seminal work in the coexpression field, "Cluster analysis and display of genome-wide expression patterns" (Eisen et al. 1998). In this work, the authors calculate the Pearson correlation coefficients between all genes for 2 different datasets. For the first dataset, they use 13 timepoints following fibroblast deprivation and reintroduction of serum, identifying 5 identifiable clusters of genes. This analysis took place prior to the completion of the human genome project, and the microarray used contained probes for only about 8600 different genes. They then extend this analysis to a fully sequenced genome, the *S. cerevisiae* genome. Collecting data from 8 different experiments that use microarrays containing every open reading frame from yeast, they calculate coexpression between all of them.

In total, these experiments contained 79 different samples from multiple different labs and conditions, providing a broad range of different yeast cell states. Following hierarchical clustering, they annotate multiple clusters with specific biochemical and homeostatic functions. These clusters included unannotated genes, highlighting the ability of coexpression to assign function to unstudied genes via guilt by association. As coexpression matured, this would become an important use case. Another important detail in this work is the early use of heatmaps for rapidly and broadly visualizing expression data. Following hierarchical clustering, gene expression patterns are clearly visible on heatmaps in a way they are not visible in tables, allowing the rapid identification of clusters.

Following this work, Hughes et al. built upon the promise of coexpression as a way to functionally annotate unknown genes. They generated 300 two-channel comparative expression profiles of *S. cerevisiae* with either mutations to a gene or chemical treatments that disturb known pathways (Hughes et al. 2000). Their goal was to establish a compendium of disturbed expression profiles that they could compare new mutants to. By identifying the treatment or known mutant with the profile most similar to the new mutant, they could characterize what pathways this known mutant was impacting. They validate this fingerprinting approach by characterizing a previously unknown yeast gene involved in ergosterol synthesis. Their gene of interest, when disrupted, has an extremely similar profile to known ergosterol synthesis genes. Testing this hypothesis, they use gas chromatography to confirm that this mutant does produce 50% less ergosterol, but because it still accumulates some it was missed by earlier screens. This key experiment showed

the ability of coexpression to identify and characterize subtle mutants that might be missed by screens or other approaches.

The next major milestone in the field was the first coexpression network built from all tissues in a multicellular organism. Stuart Kim et al. collected data from 553 two channel microarray experiments in *Caenorhabditis elegans*, using this data to build a gene expression matrix between all 17,817 genes (Kim et al. 2001). It is important to note that approximately 1/3 of these experiments used an older microarray with only about 12000 genes, slightly limiting their analysis. Instead of a hierarchical clustering approach, they used force directed placement based on each gene's top 20 most coexpressed genes, a step towards the eventual development of weighted gene coexpression network analysis. This clustering approach enabled them to plot their genes on a 2-dimensional space, with similar genes in dense clusters. This approach was able to separate genes into more distinct clusters than previous hierarchical clustering strategies, identifying 43 "mounts" of coexpression clusters. The authors also highlighted the importance of generating coexpression networks from large numbers of diverse samples, an important detail for building high quality coexpression networks. Using large numbers of diverse samples improves datasets by reducing noise and washing out the technical effects specific to a single laboratory.

Having established that large numbers of diverse samples can improve the quality of coexpression networks, meta-analysis of existing expression data became an important part of the field, empowered by two main factors. In addition

to the increasing amount of expression data generated, more publishers began requiring the deposition of generated data into standardized repositories, instead of being provided along with the paper, by request, or on personal websites. In a 2004 paper, Lee et al. used data from 60 different experiments (totaling 3924 different microarrays) to identify coexpression links within each dataset (H. K. Lee et al. 2004). These raw links, identified in single datasets, were filtered by removing any coexpression links not identified in at least 3 data sets. Surprisingly, only 2.2% of the approximately 10 million raw links were present in 3 or more datasets, and most of these genes were those with positive correlations. To confirm the hypothesis that genes with reproducible coexpression links were more likely to be real, they checked if these genes were more likely to have similar GO annotations, finding that the reproducible links did indicate functional similarity. There were two main takeaways from these results. First, that coexpression results from an individual lab's set of experiments were noisy, and that by combining many experiments across laboratories, we could increase our ability to detect "true" coexpression relationships. This type of meta-analysis is very powerful, and meta-analytic networks form the basis of much of this work. Secondly, that many negative correlation results are spurious, and we should be very careful when interpreting these negative correlations.

With the creation of coexpression networks in multiple species, it was inevitable that the next major advancement in the field would be to compare coexpression networks across these species. In 2003, Joshua Stuart and Stuart Kim would attempt this for the first time, comparing *Homo sapiens, Drosophila*

melanogaster, Caenorhabditis elegans, and Saccharomyces cerevisiae (J. M. Stuart et al. 2003). The main challenge for this cross-species comparison was identifying what genes they could compare across species. In the millions of years of evolution between these species, many genes had been duplicated or lost. To address this, they selected only genes that were the reciprocal best BLAST hit for each other, allowing them to be missing from some of the species. This identified approximately 6000 meta-genes in the multicellular species, and 2500 genes in yeast. Taking data from over 2600 experiments in total, they generated coexpression networks for each organism using these meta-genes, as well as multispecies networks. Taking these coexpression networks, they checked what percentage of different KEGG pathways were linked by these networks. They found that pathways such as ribosome and proteasome were highly covered across the multispecies and individual species networks, while cell cycle was much less conserved. Following up on the observation that older, critical functions were more deeply conserved, they investigated genes by age, finding that genes that were not animal specific were the most deeply conserved. This result is foundational to cross-species coexpression and has been replicated in dozens of species and in multiple kingdoms. Genes that are older and more vital to homeostasis, like those related to protein synthesis and DNA repair, tend to be more conserved across species.

Moving forward several years, the advent of the RNA-sequencing (RNA-seq) era in 2006 enabled more accurate and quantitative measures of mRNA expression (Bainbridge et al. 2006). However, partially due to the high initial cost

of collecting this data, it took several years before the creation of coexpression networks built from RNA-seq data and their comparison to microarray based networks. In 2012, Iancu et al. built coexpression networks from striatum samples in mice, generating a network based on microarray data and a network based on RNA-seq data (Iancu et al. 2012). Because microarrays can be truncated at the low end (not enough binding to detect fluorescence) and the high end (by saturation), they hypothesized that RNA-seq would better capture the true coexpression network. Comparing the two networks to see if the higher dynamic range and sensitivity of RNA-seq produced a network that had higher correlation coefficients, as well as more variable coefficients, they found that the higher dynamic range of RNA-seq did improve these aspects of the network. Additionally, they found that RNA-seq-based coexpression networks converge on their final network more rapidly as more samples are added than microarray-based networks. This suggests that RNA-seq more accurately captures the underlying true expression than microarrays. Overall, the study highlights the improvement in data from microarray to RNA-seq, and how important this is to producing high quality coexpression networks.

Jumping ahead only two years to the current era of single cell RNA sequencing, Biase et al. published the first coexpression network generated entirely from single cell RNA-sequencing (scRNA-seq) data (Biase, Cao, and Zhong 2014). Using single cell data from 49 individual cells, the authors generated two coexpression networks, one from 20 cells from 2-cell embryos, and one from 20 cells from 4-cell embryos. While this was the first reported network generated from

only single cell data, the sparse initial dataset, and highly homogenous starting cell populations strongly limited the broader applicability of this use case, and the authors drew few conclusions based on this network. Later work by Suonan Chen and Jessica Mar would criticize this and other early single cell coexpression work, using both real and simulated data to show that sample sizes, methods, and data quality were not good enough to build accurate coexpression networks from the data available at the time (S. Chen and Mar 2018). After several years of technological maturation enabled larger sample sizes of cells and increased sequencing depth, Xuran Wang et al. build upon earlier work by Baran developing Metacells to build a coexpression network using Metacells as the base unit, instead of individual single cells (Baran et al. 2019; Xuran Wang, Choi, and Roeder 2021). This innovation reduced zeros in the data and smoothed sampling spikes, dramatically improving the networks and enabling the creation of high quality coexpression networks from single cell data. This method was further validated in work by Benjamin Harris using high quality single cell and bulk data from BICCN, where he showed that networks generated this network were highly similar to gold standard networks generated from bulk RNA-seq data (Harris et al. 2021).

The evolution of coexpression networks has been closely intertwined with advancements in gene expression technologies. Despite this, a clear throughline between all of these works is that while improved technology can make it easier to produce high quality coexpression networks, sufficient sampling is key to the final quality of the coexpression network. With the advent of single cell technology enabling a single lab to measure expression levels in tens of thousands of cells at

once, producing a gold standard coexpression network within a single lab may finally be within reach. Having evaluated the history of coexpression network generation and comparison, we will focus on more detailed network analysis methods that can be applied following the generation of a coexpression network.

1.4 Network analysis in coexpression networks

Stepping back somewhat, we will examine important developments in coexpression network analysis, an important area that seeks to leverage generated coexpression networks to derive insights into gene function, disease, and gene regulation. Coexpression network analysis has emerged as a powerful tool in the era of high-throughput genomics, allowing researchers to make sense of vast amounts of gene expression data. These networks represent complex relationships between genes, where connections are established based on similar expression patterns across various conditions or samples. While the earliest work was mostly limited to simple hierarchical clustering to analyze the networks, later researchers noticed and leveraged the large amount of latent information in these networks to better utilize coexpression data. I will highlight several key papers that build on these coexpression networks to associate genes more deeply with upstream causes of coregulation, improve module identification and better compare coexpression networks between conditions and species.

One of the earliest papers building on coexpression analysis was work by Ihmels et al. to identify "transcription modules", context dependent sets of genes that have several benefits over the existing hierarchical clustering methods (Ihmels et al. 2002). In their method, they refine existing sets of genes to better identify both the condition that drives coregulation and other genes that belong to that transcription module. They do this by taking a set of yeast genes, and finding which conditions trigger a large shift in that gene set. Next, they examine only those experiments, and identify genes that are highly coexpressed across those conditions. These highly coexpressed genes are a transcription module. They repeat this across thousands of tens of thousands input sets to identify recurrent transcription modules. The main benefit of this approach is that unlike in hierarchical clustering, where a gene can belong to only a single cluster, a gene can belong to multiple transcription modules. This better reflects the underlying biology, where a gene may be turned on by multiple conditions or function in multiple pathways. Additionally, this approach associates a module with a specific cause/trigger, unlike in hierarchical clustering. The first and last author further build upon this approach the next year, where they leverage the identified transcription modules and a similar algorithm to enhance cross-species comparisons (Bergmann, Ihmels, and Barkai 2004). Starting with the identified transcription modules in yeast, they use BLAST to identify homologues in 5 other species (*E. coli, A. thaliana, C. elegans, D. melanogaster, and H. sapiens*). Once they have identified homologs in the other species, they remove genes that are not strongly coexpressed with the module, and add genes that are strongly coexpressed

with it, resulting in refined modules in each species. Comparing identified core regulatory modules across all 6 species, they found that higher order correlations existed between these modules and tended to be conserved. For example, ribosomal protein and secreted protein were highly correlated in all species. They also found that genes that were coexpression hubs in their networks were much more likely to be lethal when knocked out, and more likely to have a homolog in yeast than other genes. This extension of their previous work shows how critical the identification of a gene is, and how simple hierarchical clustering can miss the true, biological relationships between genes.

Another foundational paper in the coexpression field is Bin Zhang and Steve Horvath's work establishing Weighted Gene Coexpression Network Analysis, or WGCNA (B. Zhang and Horvath 2005). This work evaluates the best way to generate a coexpression network from correlation scores – is it better to use a hard threshold and binarize the network edges or to use a soft threshold with weighted edges between the nodes? As one might guess based on the name of their method, they settle on using weighted edges. They convert the similarity matrix into an adjacency matrix, by using a function that maps low values much lower, while leaving high values close to their original value. For example, a sigmoid adjacency function might reduce values of .2 to .05, keep values of .5 at .5, and convert values of .8 to .95. They suggest using either a sigmoid or power function, and conclude that the choice between the two is not important if the parameters are well chosen. Finally, they define modules based on how topologically overlapped two nodes there are – that is, how many neighbors the two nodes share and how

connected they are. In the years since proposing this method, WGCNA has become essential to unsupervised clustering, and is one of the most cited papers in the field. This clustering method is robust and reproducible and works well for identifying modules of genes in real world coexpression data.

Almost two decades later, this work would be built upon by Morabito et al, to work with two types of novel expression data. These authors updated the WGCNA framework to work with both scRNA-seq data and spatial data, calling their updated method hdWGCNA (Morabito et al. 2023). Taking the same approach as Xuran Wang et al, they merge highly similar individual cells into metacells before calculating coexpression across the cells, generating a coexpression network. For spatial data, they take a similar approach, merging spots with neighbors to create "metaspots". Following creating of the coexpression network, hdWCGNA follows a standard WGCNA pipeline, emphasizing the staying power of the WGCNA approach.

For many years, a major goal in coexpression network analysis was to find "differential coexpression", comparisons between two networks generated from different conditions in the same species that showed a perturbation, with the goal of associating this shift in the network with a downstream condition, like disease. One of the earliest papers attempting this was from Chen et al, where they attempt to associate network shifts with phenotypic change (Y. Chen et al. 2008). The authors built liver and adipose coexpression specific networks, and attempted to identify how these networks linked a known QTL region for obesity to known

metabolic genes. Identifying 19 modules within the network, they find one highly conserved network between the two tissues that links all traits to the QTL region. They identify 3 uncharacterized genes within the module and find they too are important to metabolic control. Although the authors frame the perturbation as important to finding the disturbed module, a simpler approach where they looked for genes highly coexpressed with known metabolic genes would have likely discovered the same result. Although still controversial, in the following years perturbation and differential coexpression networks failed to materialize, supporting the idea that all cells within an organism are sampling from the same general coexpression network.

Finally, a recent approach innovated by Crow et al. improved our ability to make cross-species comparisons between coexpression networks without requiring clear gene homology between all genes (Crow et al. 2022). While traditional comparisons of coexpression between two networks requires the identification of one-to-one homologs between the two species, Crow proposes the use coexpression conservation to avoid this. For each gene in species A, the top 10 coexpression partners of that gene are captured as a "fingerprint" of that gene's function, relying on guilt by association principles. These 10 genes are limited to genes that have one-to-one orthologs, enabling comparison across species. Then, for every gene in Species B, we attempt to predict its top 10 most coexpressed genes using the top 10 list from species A. This produces an AUROC for every gene in species B, representing its functional similarity to the gene in species A. This is repeated for every gene in Species A, and then this entire process is repeated

in the opposite direction to get measures for how well the top 10 coexpressed genes in Species B predict the coexpressed genes in Species A. These scores are averaged to generate the final coexpression conservation score. Testing the properties of this score, they show that genes with high sequence conservation, very old genes, and strongly predicted orthologous genes all have significantly higher coexpression conservation scores between species.

Critically, coexpression conservation does not require the identification of one-to-one orthologs between all species, instead relying on them only to "fingerprint" the gene. Additionally, it makes pairwise comparisons between all genes in two species, allowing highly broad comparisons and searches between them. Because of these two key properties, coexpression conservation is uniquely suited towards making comparisons between plant species, where frequent gene duplications have vastly expanded plant families, and homology is unclear. When combined with the inherent, tissue and organ homology neutral properties of coexpression, this metric can make comparisons between species with minimal hand annotation of data, if sufficient RNA-seq data is available to generate coexpression networks from. In this thesis, I will build upon this method, highlighting an infrastructure leveraging it to improve the integration of single cell data and building upon its observations to better compare distantly related plant species.

Chapter 2

EPIPHITES: Enhancing cross-species integration of scRNA-seq data

In this chapter, I will introduce our manuscript in which we propose a method for improving the alignment and integration of cross-species scRNA data in plants. This is a frequent pain point during the analysis of cross-species data, and improper alignment, integration, and batch correction can lead to entirely spurious results. As more and more scRNA data is generated from novel species, accurate and appropriate alignment is critical to properly interpret it. Here, we introduce a method that leverages gold standard bulk RNA-seq based networks to expand the one-to-one gene space and show that it improves the integration of cross-species scRNA-seq data. Additionally, it works on any high dimensional genomic data, and slots upstream of most integration and batch correction methods, improving a broad range of workflows. Finally, we include a workflow for generating new coexpression networks from any single cell experiment.

2.1 Preamble

Our method, EPIPHITES (Expression Proxies In Plants Help Integrate Transcribed Expression in Single-cell), uses bulk RNA-sequencing data to identify cross-species gene pairs (coexpression proxies) that can be added to the existing one-to-one gene space when integrating data, and by doing this, increases the accuracy of integrations. Before diving into the results of this approach, I will lay out the rationale for this somewhat unusual strategy, which explicitly determines one-to-one pairs where there may be no true pairs.

Genomics data, like RNA-seq, ATAC-seq, or methylome data, is very high dimensional. A typical eukaryotic genome has on the order of 20,000 genes, and methods that also generate data for non-coding regions of the genome have even higher dimensions. Yet, we are able to accurately capture information about these datasets with much smaller sampling regimes than you would expect naively. If all genes carried equal amounts of *independent* information about a state change, differential expression analysis between two very different states could return only a single significant gene. A single gene being expressed at different between two cell types could indicate a massive difference in the state of the cell. In this scenario, we would require extremely deep sequencing to sensitively capture gene readouts for all ~20,000 genes. However, we know that this is not the case in biological data. Instead of being independent, gene expression values are highly correlated within one another, as genes are co-regulated in modules to perform key

functions. As typically one gene is not enough to accomplish a function, the coexpression of multiple genes is required for cell function. Because of this codependency, a small shift in cell state results in many coregulated genes changing their expression level, allowing us to accurately capture cell state using much lower sequencing depths than would otherwise be possible, enabling things like differential expression and GO term analysis of gene sets.

Single-cell RNA sequencing is highly dependent on this phenomenon to enable the annotation of cells. Currently, scRNA-seq is noisy and does not deeply sample the all mRNA in a cell, which is already a limited population, often capturing only single digit number of reads per gene. Despite this, we are able to cluster cells by their cell type, even when some are missing expression values for large percentages of their genes or markers. This is possible because of the high amounts of information about other genes provided by the expression of a single gene. Even if the expression overlap between two cells is low, we can cluster them together if we know that those non overlapping genes tend to be expressed with a common set of overlapping genes. Without this effect, it would be impossible to cluster single cell data at the current sequencing depth.

When integrating high-dimensional data across species, current methods require one-to-one alignment of genes across the two species, and typically drop all genes that do not have a one-to-one match. While the goal for integration of cross-species data is often ambiguous, a common goal is to minimize technical batch effects while not fitting away real, biological variation between the two

species. This is challenging, as the technical batch effects are perfectly confounded by species. For example, if one species requires a more intense treatment to remove the cell wall, this treatment will be confounded with the species, and will be difficult to tease out via integration. Success in integration is typically measured by accurate alignment of homologous cell types across species. In a good integration, homologous cell types across species should be more similar to each other than different cell types from the same species.

While there are sufficient one-to-one gene matches in mammals due to the lack of major genomic shocks like polyploidization events, in plants one-to-one gene mapping is much more difficult. Distantly related species, like maize and Arabidopsis, may share as few as 3,000 one-to-one gene pairs out of the ~20,000 Arabidopsis genes and ~30,000 maize genes. Additionally, these remaining one-to-one gene pairs often provide even less cell type and cell state information than expected, as they are often globally expressed genes key to cell survival, such as ribosomal and DNA repair genes. Because of this issue, integrating high-dimensional plant data is very challenging, as you have already lost much of your information when limited to one-to-one genes, and your remaining genes are less informative for determining cell type due to their ubiquitous expression. A final barrier to integration is the lower state of overall differentiation in plant cells. In mammals most tissues are composed of terminally differentiated cells incapable of regeneration, and a set body plan limits growth. In contrast, whole plants are typically easily regenerated from single organs, like leaves, and there is no set body

plan. Because of this, plant cell types are often less distinct from each other than mammal cell types, which makes integration, and its evaluation, more challenging.

Despite the difficulty of cross-species integration, it is a key frontier in the field of plant science. Cross-species integration is critical to directly compare cell types between species and understand what modifications have occurred as the two species evolved. Identifying changes in gene expression and regulation is important to understanding the mechanisms that underly key traits like drought tolerance that have evolved in some species and are desirable in crop species. However, direct comparisons in the field have been held back by the difficulty of accurately integrating distant species. Previous work in cross-species scRNA-seq has focused on making comparisons between closely related species, limiting comparisons within families. We sought to improve these integrations by expanding the one-to-one gene space, the main barrier to integration.

Fortunately for us, decades of work have been done to track the evolution of genes between species, establishing the ancestry of genes and relating them to one another in and across species. These species represent groups of genes that descended from a common ancestor defined at a certain age, and thus can be expected to have very similar functions. We use these gene families as the starting point for improving the one-to-one gene space, limiting our search space to reduce the total comparisons made, and ensuring that we select genes that are at minimum reasonable matches for one another. Next, using existing gold standard bulk coexpression networks we determine the conservation of coexpression between all

possible pairs in the family, selecting only reciprocal best hits that are substantially better than other options. We also discard real one-to-one pairs that have extremely low performance. We name these identified additional one-to-one pairs coexpression proxies.

The rationale behind this strategy is that because we are integrating a very high dimensional space, we do not need to get every additional coexpression proxy correct. A large issue in existing integration pipelines is the lack of one-to-one genes, severely limiting the data set. If we get most of our coexpression proxy calls right, and if the rest are close to right (genes with highly similar if not exact expression profiles), we will improve the integration. By adding enough coexpression proxies that are close to correct, we can massively improve the integration as incorrect calls are washed out by good matches. To give a toy example, imagine a gene family with 8 genes in Species A and 12 in Species B. Normally, we would have no gene pairs from this family, as there are no one-to-one matches. By using our approach of matching genes based on conservation of coexpression, we might successfully call 5 pairs of coexpression proxies. If 4 of these are high quality matches that improve the integration, and 1 is a low-quality match, we would expect overall integration quality to improve, as signal from the 4 genes washes out issues from the single bad call.

Additionally, by calling gene pairs, we can improve the panoply of existing integration methods, as these methods require one-to-one matches for integration. This allows our method to be upstream of existing methods, improving them all

and allowing robustness testing between multiple integration strategies, a critical part of evaluating an integration. Because a substantial amount of effort has been put into developing high quality integration methods for non-cross-species integrations, an approach that feeds into these methods instead of creating a separate workflow enables future authors to have more flexibility in what integration approach they take.

Overall, we take a simple approach that leans heavily on well validated, robust coexpression networks derived from bulk RNA-seq data. We use these networks to identify genes with highly similar expression profiles between pairs of species, expanding the shared gene space and feeding into existing integration methods. Below is our manuscript highlighting the success of this approach to improving integration.

2.2 Coexpression enhances cross-species integration of single-cell RNA sequencing across diverse plant species

2.2.1 Abstract

Single-cell RNA sequencing is increasingly used to investigate cross-species differences driven by gene expression and cell-type composition in plants. However, the frequent expansion of plant gene families due to whole genome duplications makes identification of one-to-one orthologs difficult, complicating

integration. Here, we demonstrate that coexpression can be used to trim many-to-many orthology families down to identify one-to-one gene pairs with proxy expression profiles, improving the performance of traditional integration methods and reducing barriers to integration across a diverse array of plant species.

2.2.2 Main Text

Plants have a remarkably flexible cellular physiology, driving their adaptation into nearly every environment. Recently, the advent of single-cell RNA sequencing (scRNA-seq) has provided novel insights into the diversity of cell types underlying these adaptations (Denyer et al. 2019; Ryu et al. 2019). The unique diversity in plants makes comparative assessments between species important but is also complicated by uncertain homology relationships. Unlike in mammals, where homologous genes and structures can be easily identified, plant gene families frequently expand by whole genome duplication, polyploidization and tandem gene duplication (Su et al. 2002; Gharib and Robinson-Rechavi 2011; Clark and Donoghue 2018). This scarcity of one-to-one gene pairs is a major barrier to defining a common gene space for the integration of single-cell data, a key step for successful cross-species comparative analysis or integration (Bennetzen 2000; T. Stuart et al. 2019). With vast amounts of plant scRNA-seq data becoming available (H. Chen et al. 2021), this study aims to address a critical gap in its analysis by using coexpression to identify pairs of genes that, while not

exclusive orthologs, are functionally related enough to enable the integration of this high-dimensional data. By reducing barriers to integration, we prime the field for the discovery of novel, cell-type specific innovations that have been critical to plant adaptation and domestication.

While a given plant sample may have thousands of expressed genes, the expression patterns of these genes are not independent, and are instead organized into the regulatory programs which underlie cell types. This coexpression generates the low-dimensional expression space that is foundational to the success of modern single-cell analysis (Heimberg et al. 2016). We hypothesize that genes with highly similar expression profiles between two species can be used as reasonable proxies for integrating cell-type specific data, that we can identify such profiles using coexpression, and that this will expand the shared gene space, improving our ability to compare cross-species data. The essence of the approach is to use meta-analysis from prior bulk RNA sequencing data to define cross-species gene pairs (coexpression proxies) that can be applied in more specific, but sparser, single cell data. By utilizing robust coexpression networks built from over 16,000 publicly available RNA sequencing datasets, as well as gene phylogenies from OrthoDB, we ensure that the coexpression proxies accurately reflect the underlying biology of each species pair they are drawn from (J. Lee et al. 2020; Kriventseva et al. 2019). We illustrate this approach, where coexpression data and gene phylogenies identify gene pairs that expand the one-to-one gene space, improving data integration and alignment between known cell types and

highlighting novel ones between species (Figure 1A). While previous work has expanded the shared gene space by gene homology comparisons, our focus on coexpression uniquely captures both regulatory and functional shifts between species (Tarashansky et al. 2021). By improving the integration, we enable authors to identify new and conserved cell types in their scRNA-seq data. We validate the coexpression proxies with two test examples, highlighting their utility. In the first test, we show that coexpression proxies can accurately reintegrate a split dataset with no shared gene space. Second, we show that coexpression proxies improve the integration of real single-cell data between two species with complex genomes: maize and rice.

Our first test is an extreme one in which we generate and integrate two cross-species datasets with no one-to-one orthologs. This would be impossible with a traditional integration approach, which requires directly matched one-to-one orthology relationships between genes in each species for alignment prior to integration. To construct a case with a ground truth integration without using synthetic data, we split an existing Arabidopsis single cell dataset into two pseudo-"species". The first "species" is generated by randomly selecting half of the cells as well as half the genome. For these cells, the second half of the genome is removed. We then take the remaining cells, which will become our second "species", and remove the half of the genome present in the first set of cells (Supplementary Figure 1). This provides two sets of cells with known, shared cell types and distinct genomes. We then identify coexpression proxies between the

two subset genomes, finding pairs of genes with similar expression profiles. As an example, the selected coexpression proxy gene, *AT1G16150*, closely matches the expression profile of the target gene, *AT1G16160*. In contrast, *AT4G31100*, a rejected gene from the same ortholog family, has a distinct expression profile (Figure 1B).

Next, we used these coexpression proxies to reintegrate the split Arabidopsis dataset. Highlighting that coexpression proxies smoothly integrate into existing workflows, we used Scanorama (Hie, Bryson, and Berger 2019) to reintegrate and re-cluster the dataset, placing 82% of cells into a cluster with cells from both datasets (Figure 1C). Additionally, the reintegration was accurate, successfully matching cells of the same cell type across datasets 75% of the time. To evaluate how much of the gene proxies' success is dependent on information from the gene phylogenies, and how much information is derived from the coexpression conservation profile, we attempted to integrate the datasets using the worst rejected proxy from within each ortholog group (i.e., lowest coexpression). Performance is lower using these gene pairs, reducing the successful matching of cells to 65% (Figure 1D). This moderate performance suggests that simple relaxation of orthology constraints is a substantial contributor to performance. However, coexpression provides a significant overall signal boost. This was particularly clear for phloem, which was otherwise unintegrated or mixed with atrichoblasts and xylem. To determine whether sequence similarity alone would prove sufficient, we calculated the pairwise protein sequence similarity of every

Arabidopsis gene and attempted to use this to identify gene proxies. While able to perform better than random, this metric was worse than coexpression at reintegrating the split dataset and completely failed to reintegrate certain clusters. Finally, we attempted integration using 1900 random gene pairs and find that we are unable to achieve any integration (Figure 1E). To further evaluate our coexpression proxies, we assess the degree to which rejected and selected gene pairs show the same expression across cell-types on a per-gene basis (measured by Euclidean distance). We find that accepted coexpression proxies are much closer to the target's expression profile across cell types, and that the rejected proxies are on average 83% further from the target's expression (Figure 1F). This shows that the coexpression proxies are more similar in expression profile to their target genes than even other genes from the same orthogroup.

Given the success of our approach, we generated coexpression proxies between 13 plant species and identified an average of 5,750 gene pairs between species (Figure 1G). The coexpression proxies are numerous enough to provide additional information across even highly diverged species and are well represented (4,899 pairs) even between *Zea mays* and *Arabidopsis thaliana*, which diverged 160 MYA. Importantly, while we used Scanorama, these coexpression proxies can be easily incorporated into any potential integration pipeline as they simply expand the shared feature space.

Having shown that coexpression proxies could integrate an otherwise uncorrectable dataset, we tested their ability to improve the integration of single

cell data across two different species. Using a supervised integration, we attempted the integration of two root datasets, one from maize and one from rice. We focused on 4 broad cell types for which author annotations directly aligned. Using coexpression proxies, we successfully integrated the maize and rice dataset, accurately integrating 36% of cells into clusters with cells from both datasets (Figure 2A, Supplementary Figure 2). The remaining cells were different enough to still appear as distinct sub-clusters across species. While this is far from 100%, real cross-species differences do exist, so it is not clear what the maximum plausible percentage is. Importantly, our integration is better than using only the one-to-one gene pairs, which integrated only 14% of the cells (Figure 2B). Key cell-types, such as epidermis and stele, are well integrated using coexpression and are less well integrated by one-to-one gene pairs, as evidenced by lack of species mixing within cell-types and close proximity across cell-types. Similarly, coexpression did not overfit away real differences, capturing the likely real difference between cortex cells where constitutive aerenchyma formation is critical to oxygen diffusion in partially submerged rice (Colmer and Pedersen 2008). To evaluate the integration on a cell-type by cell-type basis, we employed MetaNeighbor, which enables us to quantify the degree to which cell types replicate across datasets in a statistical framework (Crow et al. 2018; Fischer et al. 2021). We compare 4 integrations using scGen — utilizing coexpression proxies and one-to-one genes, only coexpression proxies, only one-to-one genes, and using random genes (Figure 2C). As can be seen, coexpression proxies alone, one-to-one pairs alone, and the combination all accurately and similarly group cell types

across species. While subtle for this broad classification, the full coexpression proxy set integrates better than either of its parts in all cell types when evaluated by MetaNeighbor (except cortex, where all methods are perfect), reflecting the additional information from the coexpression proxies. Because this is a validation focused on well-defined alignment, performances generally go from high to even higher (e.g., stele goes from AUROC 0.93 to 0.973). To evaluate utility of an increased known gene-pair space, as well as the robustness of the model, we swapped in coexpression proxies for random pairs and tracked performance improvement (Figure 2D). Performance increases steadily to near 1 for most cell types, indicating that the typical number of 5000 coexpression proxies is sufficient to integrate cross-species data. Further querying the coexpression proxies, we found they typically represented core conserved functions such as photosynthesis, mitochondrial proteins, and ribosome metabolism (Figure 2E).

Integrating cross-species single-cell data is an increasingly common goal in the fields of plant development, evolution, and molecular biology. To facilitate this process, we have demonstrated that using coexpression proxies expands the gene space available for integration. To facilitate adoption of this approach by the community, we have generated pairwise coexpression proxies between 13 plant species at 3 thresholds. All coexpression proxy lists are made available at https://gillislab.shinyapps.io/epiphites_v11/. Additionally, we have provided a workflow for generating a coexpression network from scRNA-seq data and using it to identify coexpression proxies for integration (Supplementary File 1), which

requires only gene phylogenies between the two species. We show that this approach generates networks similar to gold standard networks and enables similar integration (Supplementary Figure 3, Supplementary Figure 4). These proxy lists provide an important resource for improving the integration of single-cell data, accelerating the transfer of knowledge from well-studied model organisms to crop systems that are crucial to the global food supply.

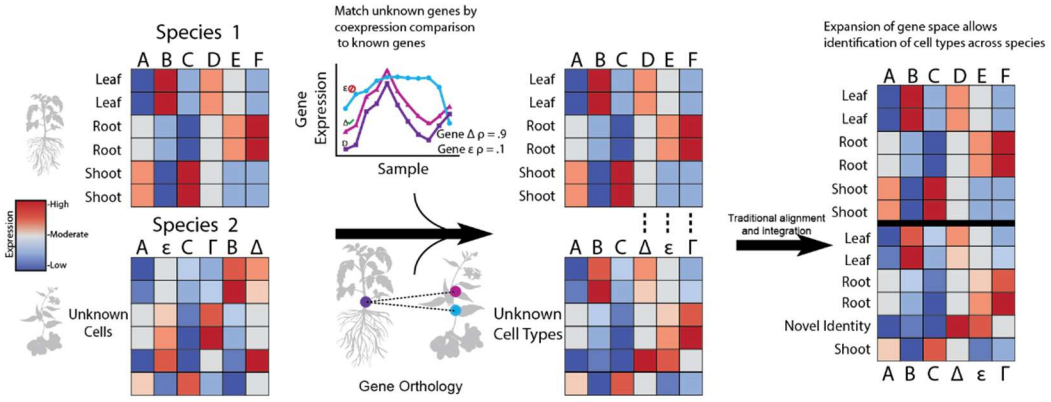


Figure 1A. Schematic depicting the identification of coexpression proxies from gene orthology information and their use in expanding the gene space to enable integration followed by identification of novel and conserved cell types

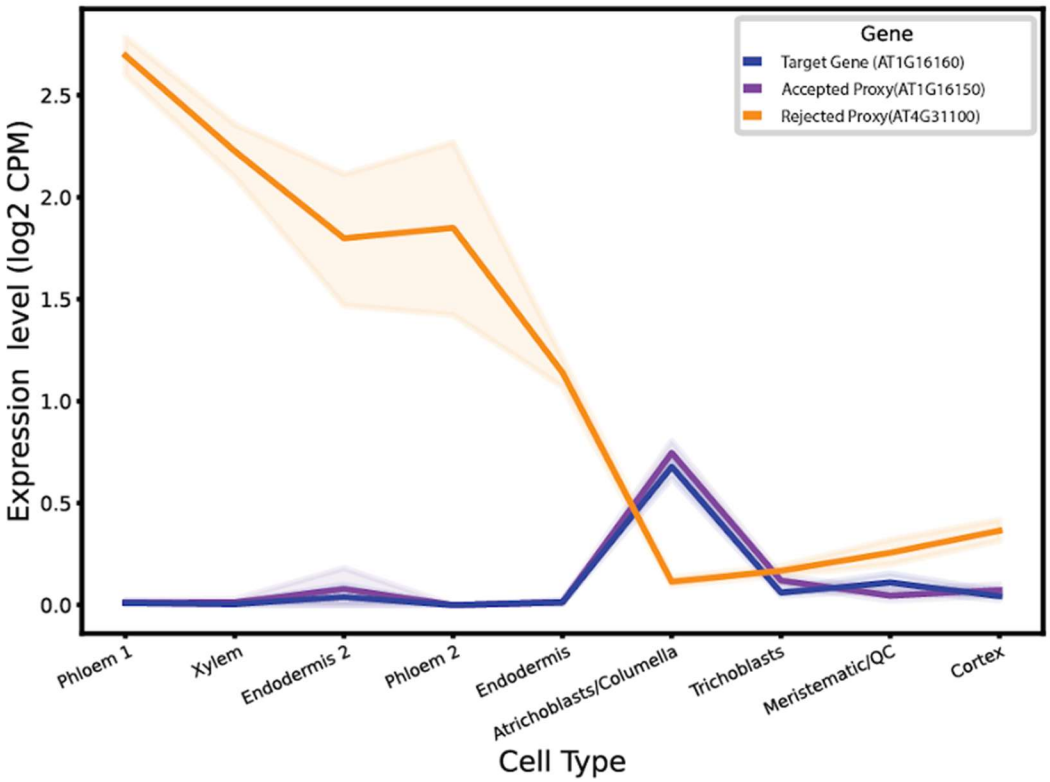


Figure 1B. Gene expression profile for target gene (AT1G16160) and two potential coexpression proxies (AT1G16150, AT4G31100). The gene with the more similar profile, AT1G16150, was identified as a coexpression proxy, while AT4G31100 was rejected. The center band is the mean counts per million (CPM) for each gene in the cell type in our single-cell dataset. The error bar is the 95% confidence interval. QC, quiescent center.

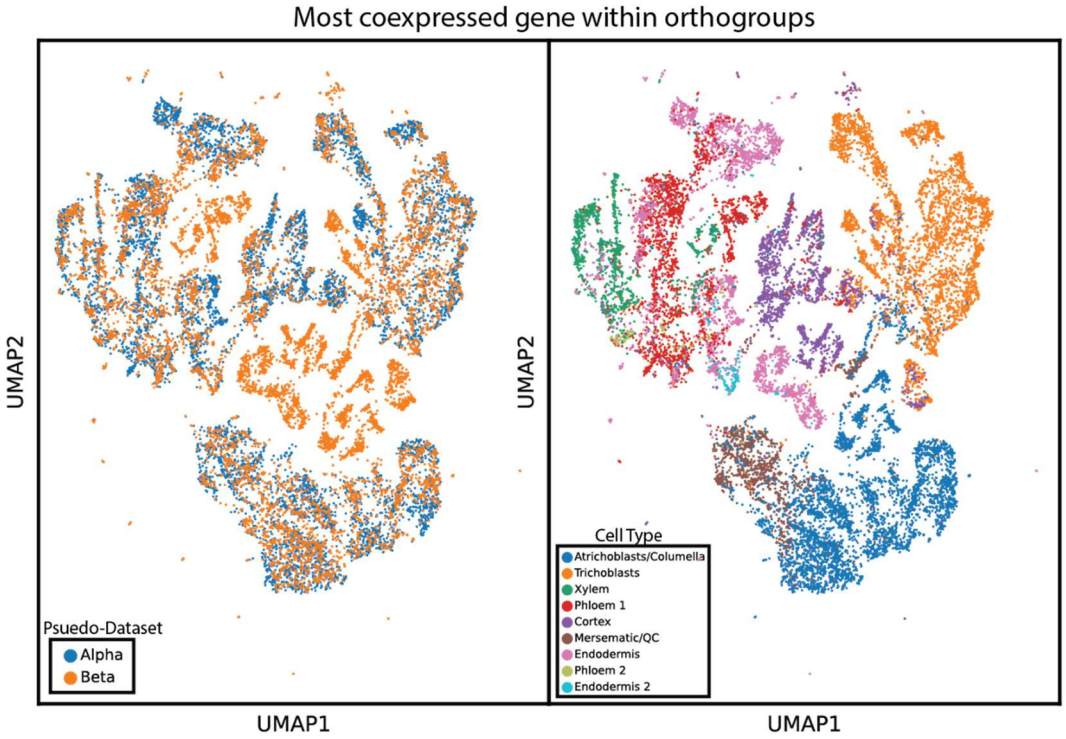


Figure 1C. UMAP showing integration of a cleaved and dissociated A. thaliana dataset containing 16,636 cells using coexpression proxies.

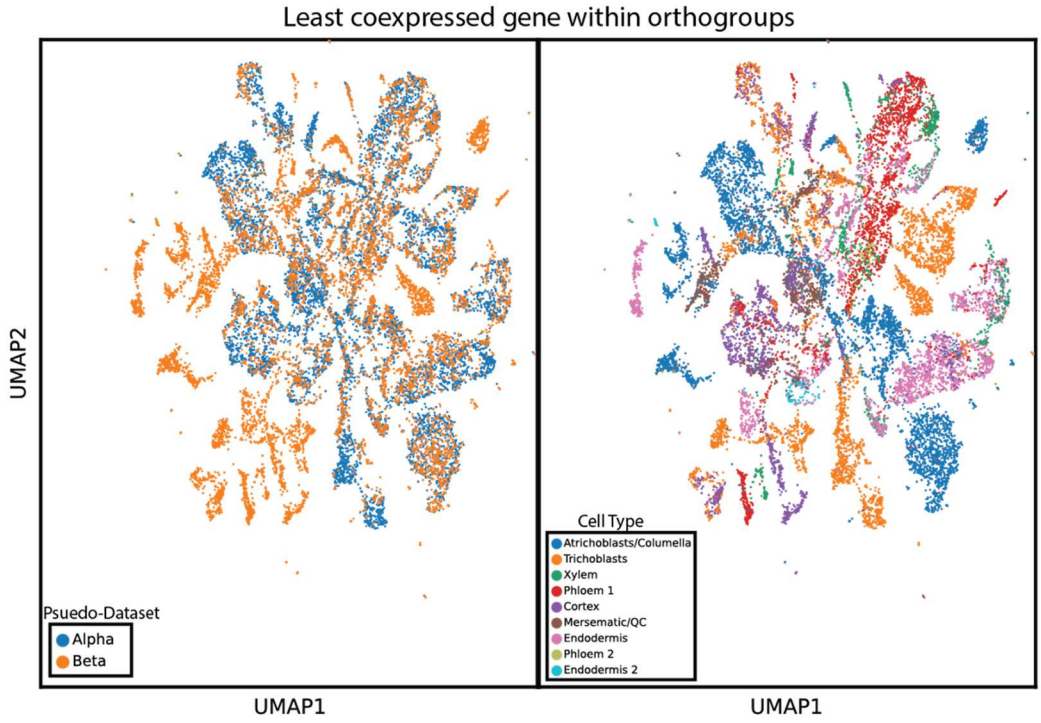


Figure 1D. UMAP showing integration of the same dataset using the worst potential coexpression proxy from each gene family.

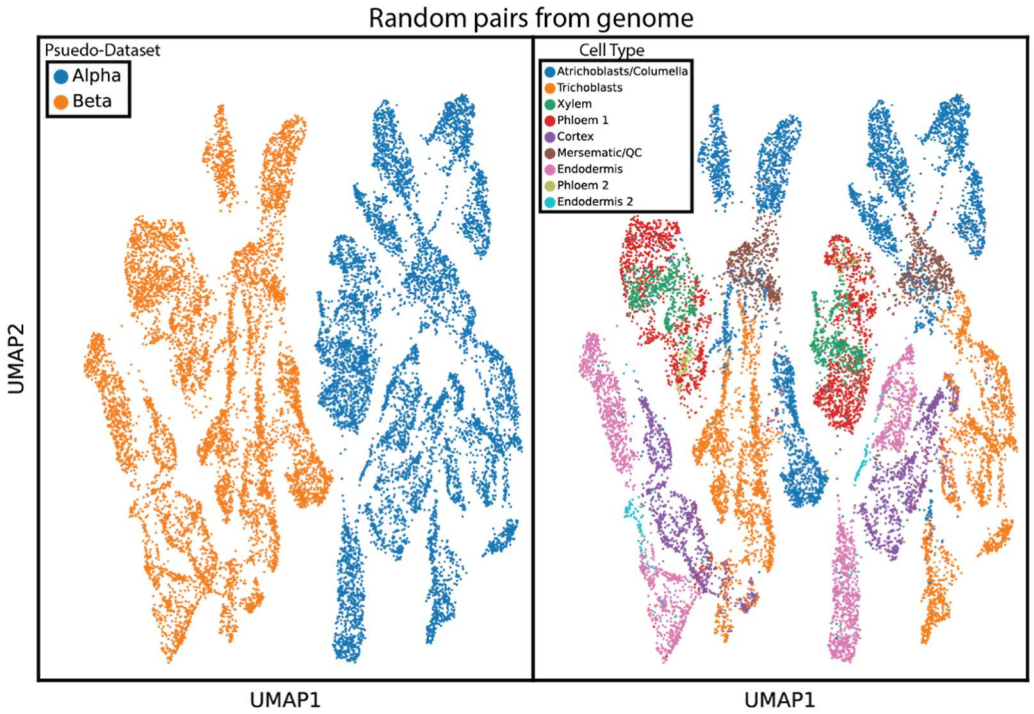
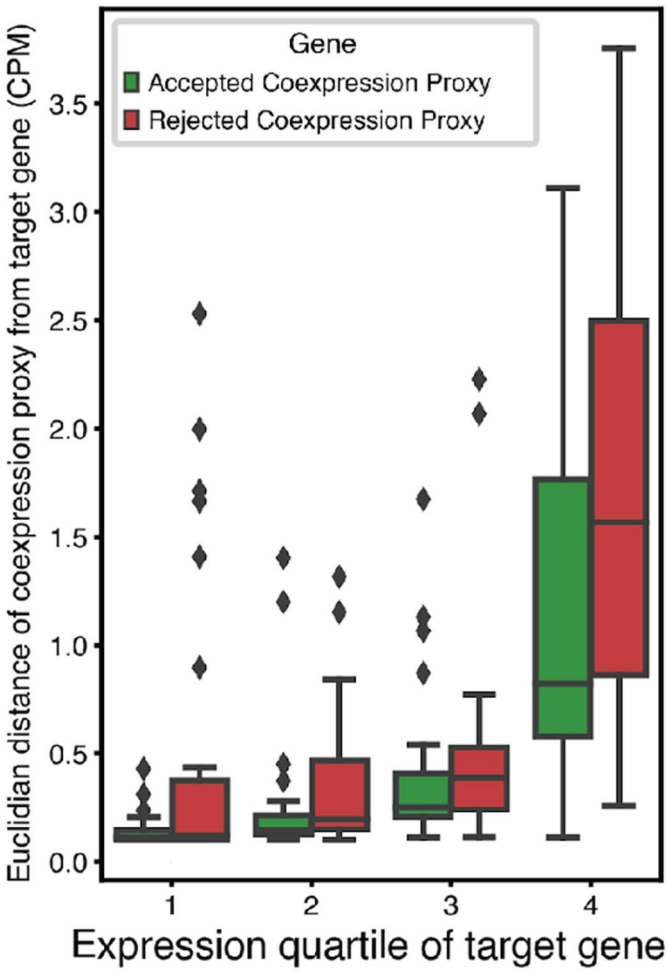


Figure 1E. UMAP showing the failed integration of the split and dissociated dataset using 1,900 random gene pairs.


Gene
Accepted Coexpression Proxy
Rejected Coexpression Proxy
Euclidian distance of coexpression proxy from target gene (CPM)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1
2
3
4
Expression quartile of target gene

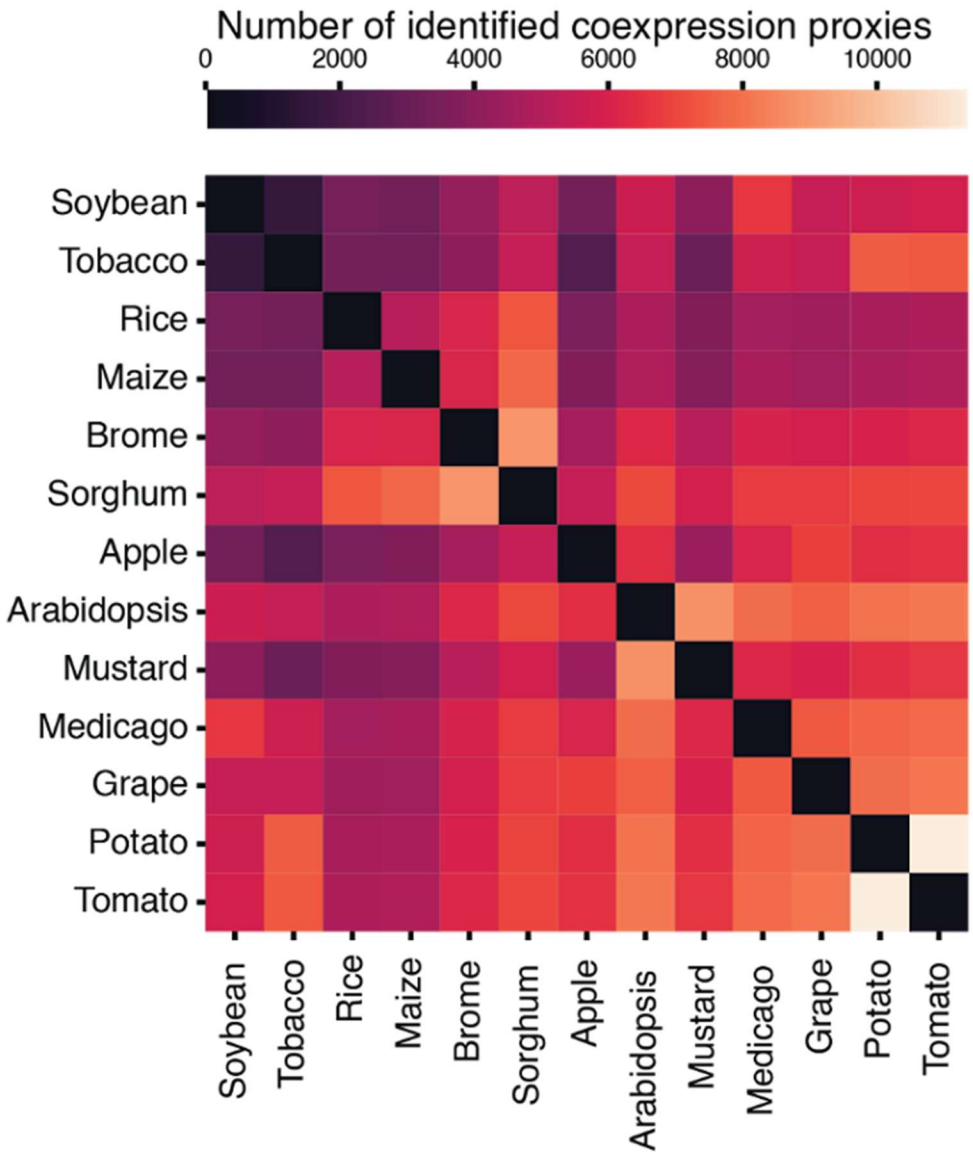


Figure 1G. Heat map showing the number of identified coexpression proxies between each species pair in the database.

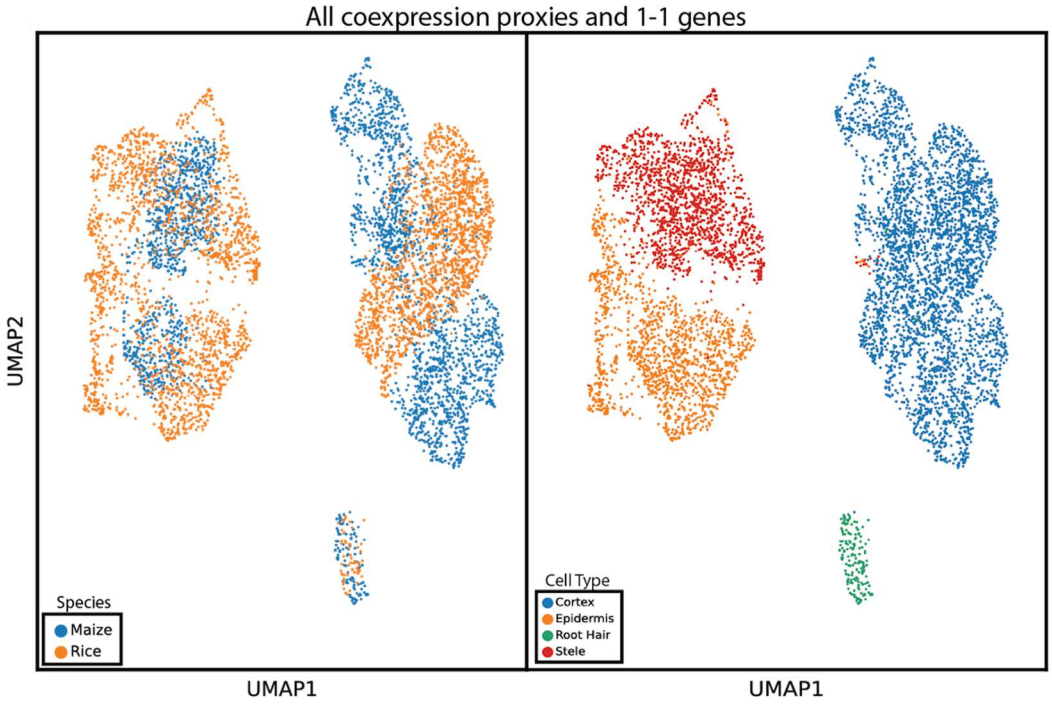


Figure 2A. UMAP showing integration of 2,832 Z. mays cells and 3,500 O. sativa cells using coexpression proxies.

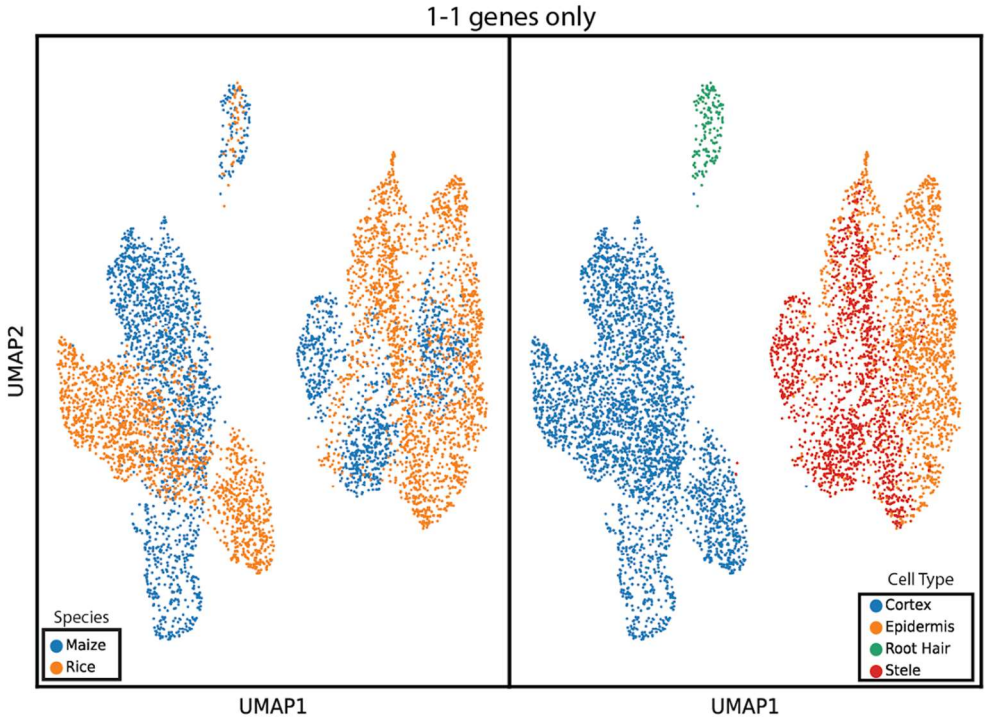


Figure 2B. UMAP showing integration of Z. mays and O. sativa using only one-to-one gene pairs from OrthoDB.

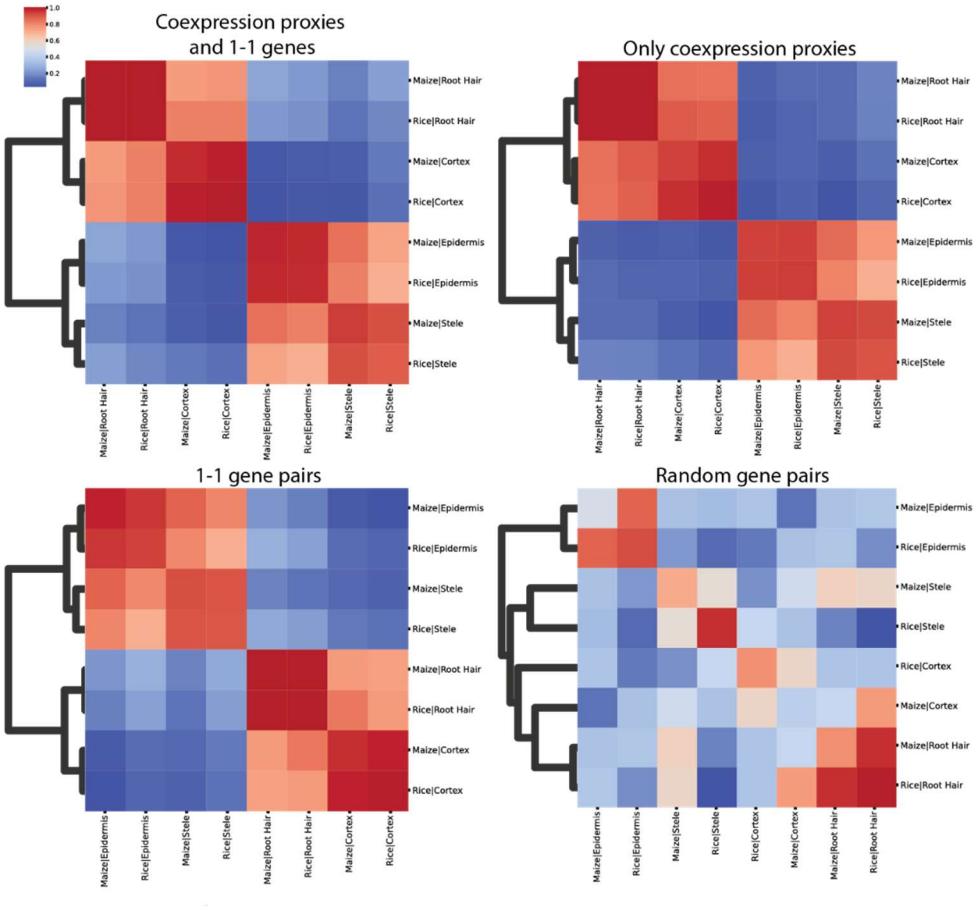


Figure 2C. MetaNeighbor plots showing post integration similarity between cell types using four different gene sets.

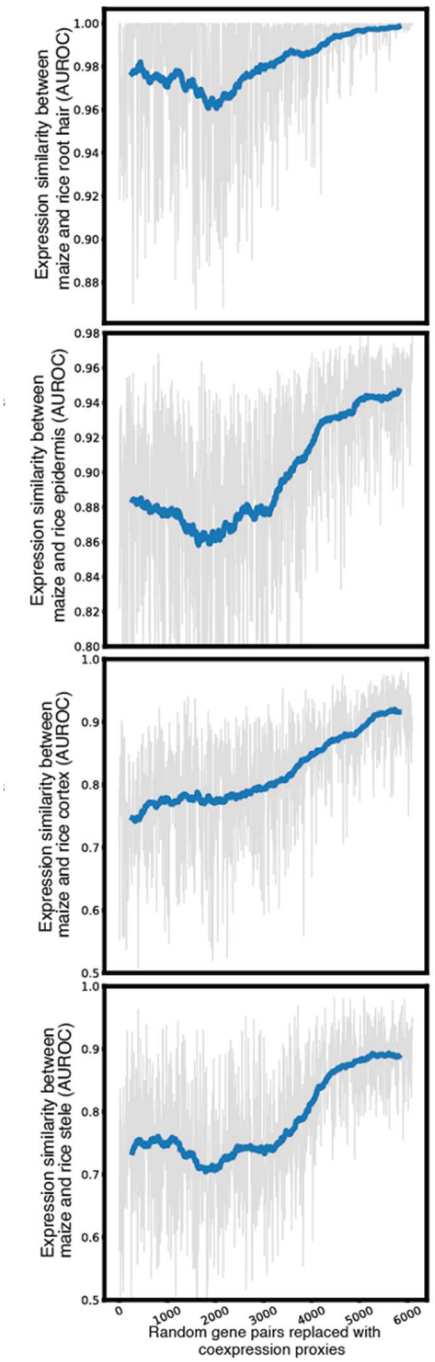
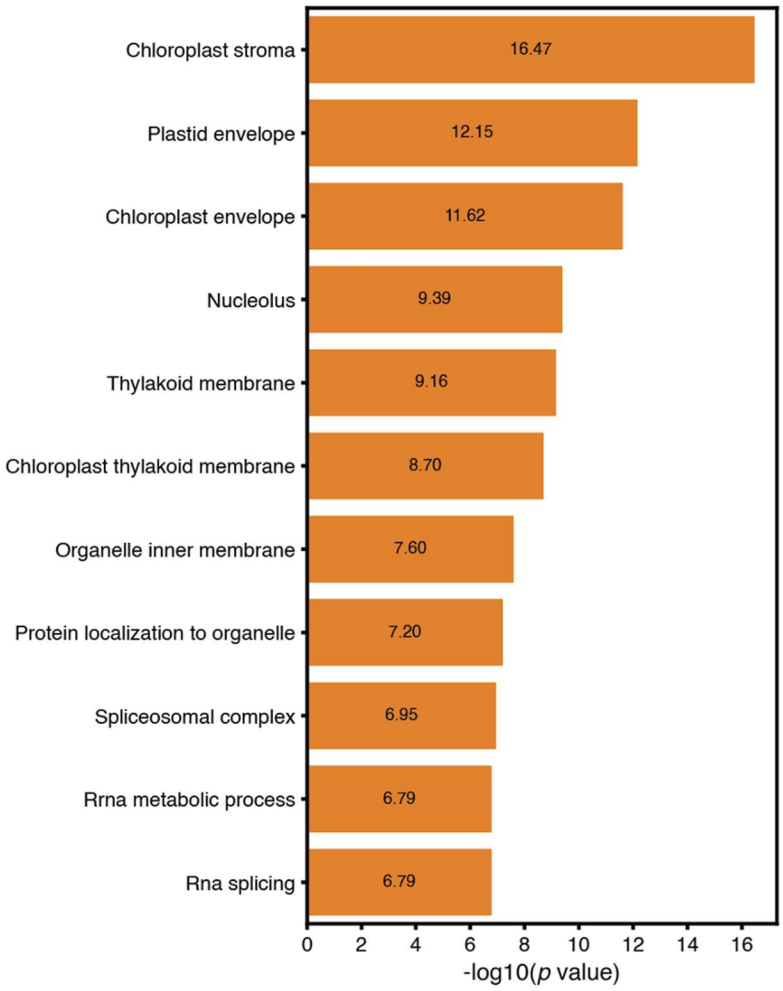
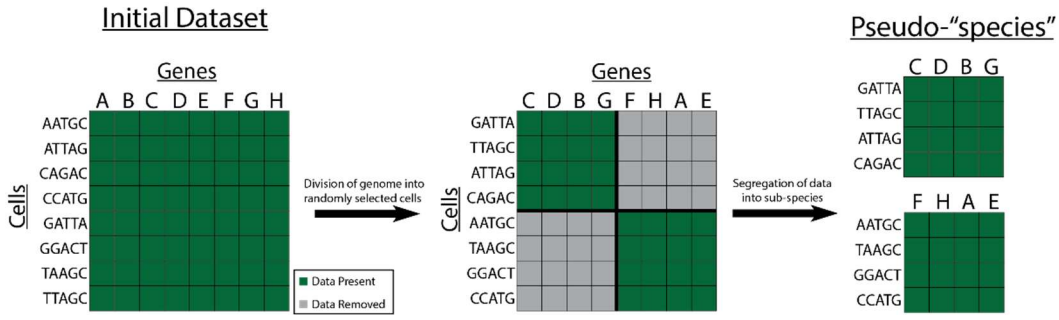
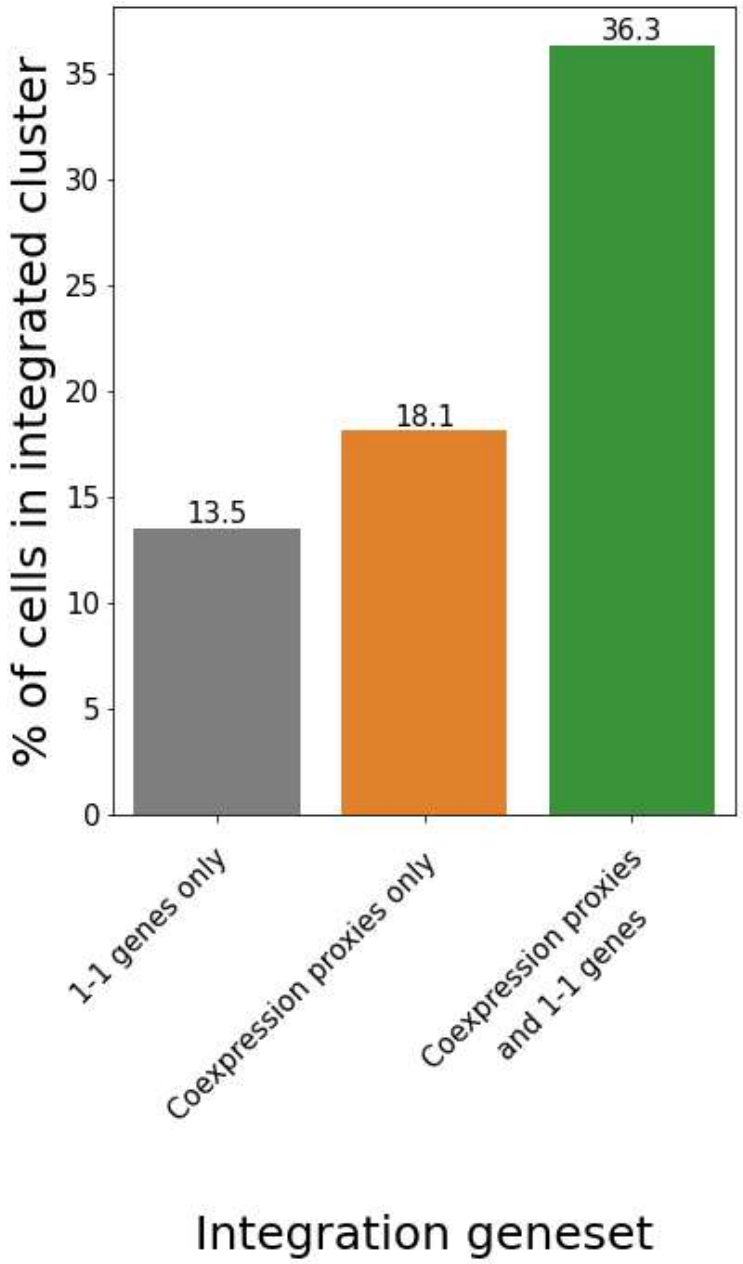


Figure 2D. Improvement in integration across 872 integration runs as random gene pairs are gradually swapped for coexpression proxies.

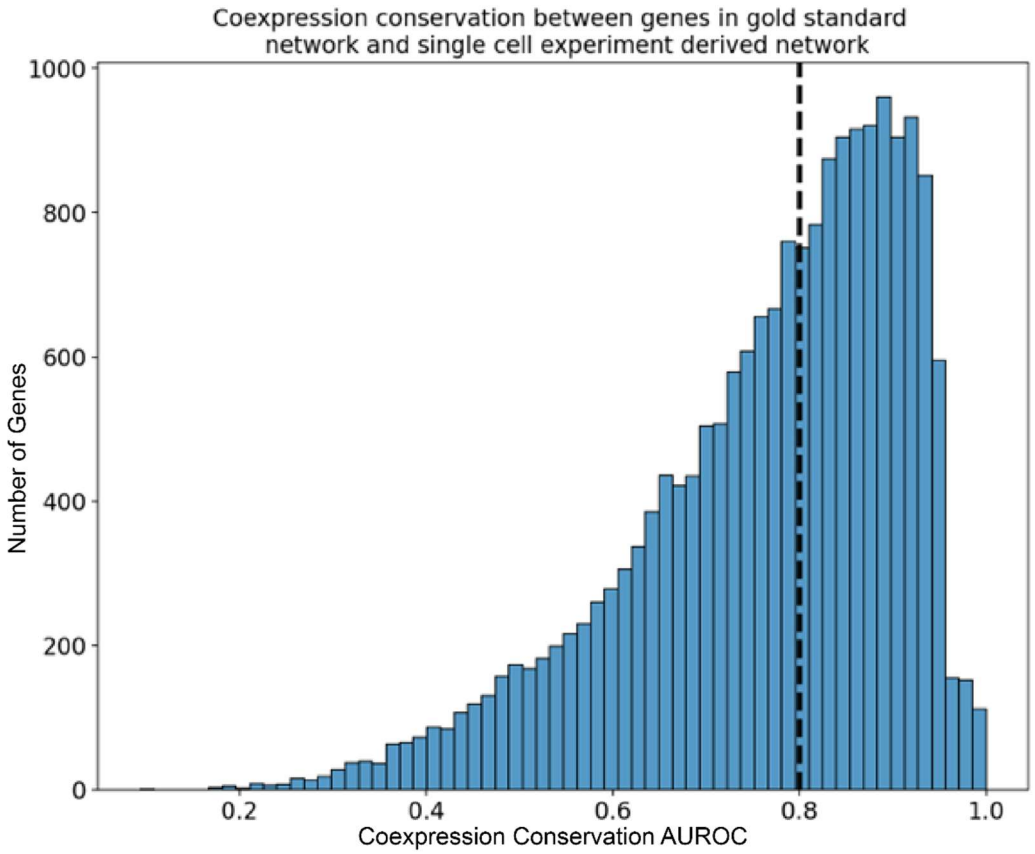

Chloroplast stroma
16.47
Plastid envelope
12.15
Chloroplast envelope
11.62
Nucleolus
9.39
Thylakoid membrane
9.16
Chloroplast thylakoid membrane
8.70
Organelle inner membrane
7.60
Protein localization to organelle
7.20
Spliceosomal complex
6.95
Rrna metabolic process
6.79
Rna splicing
6.79
0
2
4
6
8
10
12
14
16
-log10(p value)



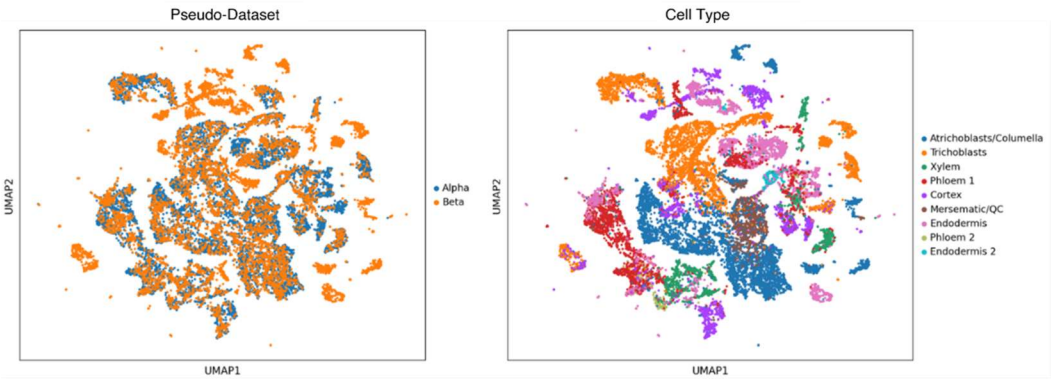
Supplemental Figure 1. Schematic illustrating how an initial dataset consisting of cells from one species is split into two datasets with no shared genes or cells, generating two "species" with a ground truth of shared cell types.



Supplemental Figure 2. Bar chart showing the percentage of maize and rice cells in a cluster with either dataset comprising no more that 70% of the cluster.



Supplemental Figure 3. Distribution of coexpression conservation scores for every Arabidopsis thaliana gene between the existing gold standard network and a new coexpression network generated using our workflow on Arabidopsis thaliana root cell data.



Supplemental figure 4. UMAP showing integration of a split and disassociated Arabidopsis thaliana dataset containing 16636 cells. The alpha dataset was assigned the existing gold standard coexpression network, and the beta dataset was assigned a new coexpression network built using our workflow for scRNA-seq data, and 1293 coexpression proxies were identified between the datasets for integration.

2.2.3 Methods

Gene coexpression proxy identification

For each species pair, gene family orthology information was downloaded from OrthoDB V11(Kriventseva et al. 2019). Utilizing one to one gene pairs, coexpression conservation was calculated between all genes in each species (Crow et al. 2022). Briefly, we compare each gene's top 10 coexpression partners across species. These top 10 are limited to genes that are one-to-one orthologs, although the matching of proxies is not limited in this way. Using one-to-ones as a basis set for comparison of other genes expands the range of potential proxies while still leaving it grounded in defined cross-species overlaps. We use the ranks from one species to predict the coexpression partners of the second species, and then repeat this in the other direction, averaging the scores to generate the conservation of coexpression score, which is an area under the receiver operator curve (AUROC). This resulted in a Species A `genes by Species B genes matrix, filled with the AUROC score for each gene pair. For each gene family, the coexpression conservation matrix was filtered to every possible cross-species gene pair. Next, pairs in multigene groups were eliminated by thresholding in two steps. First, any gene pairs with scores below a quality threshold were discarded. Second, remaining pairs were required to be reciprocal best hits, and to be higher than other potential options by a multi-pair threshold. For genes that were one to one matches,

they were only discarded if below a lower single pair quality threshold. For the moderate filtering, the quality, multi-pair and single pair junk thresholds were 0.85, 0.03, and 0.8. For lenient and stringent filtering, the thresholds were 0.8, 0.02, 0.07, and 0.9, 0.035 ,0.085, respectively. The moderate threshold was chosen by evaluating the number of proxies identified at many thresholds and choosing the elbow, and lenient and stringent thresholds were picked to form a 0.1 range around this number.

Dataset integration and evaluation

To generate an integration task that was uncorrectable without a shared gene space, the Arabidopsis dataset was split into two sets of cells. Using Pandas DataFrame.sample, one half of the genome was randomly selected and assigned to the first set of cells, with other data being discarded. The second set of cells were assigned the second half of the genome, and the genes assigned to the first half were discarded. Utilizing the same method as above, coexpression proxies were identified between the two halves of the genome with the moderate threshold. Aligning the two gene spaces using these proxies, we performed integration using the Scanorama Python package (Hie, Bryson, and Berger 2019). The scanorma.integrate function was used to integrate the two datasets into a shared low dimensional space, and this was plotted using scanpy.pp.neighbors with the default parameters (15 nearest neighbors, 50 PCs), and the scanpy.tl.umap function (default parameters). For evaluation, we first clustered the integrated data using scanpy.tl.leiden at a resolution of 0.5. This provided an evaluation space that is

based on the high dimensional underlying data, instead of the 2D UMAP projection, which can be misleading. Then, using these clusters, we defined a cluster of the same cell type as one containing more that 60% of that cell type, and a mixed cluster as one composed of between 30-70% of each starting dataset. In plotted boxplots, center line is the median, the box limits are the upper and lower quartiles, the whiskers are 1.5x the interquartile range, and the points are any datapoints beyond the whisker range.

As the cross-species integration scenario was more challenging, it was integrated utilizing scGEN version 2.1.0(Lotfollahi, Wolf, and Theis 2019). The datasets were limited to 4 broad cell types for which author annotations clearly aligned, and the rice dataset was subset to 3500 cells to match the maize dataset of 2832 cells. The tissues were equally represented in each of the two datasets. Utilizing coexpression proxies between rice and maize at the moderate threshold, the two datasets were aligned. The two datasets were first aligned utilizing coexpression proxies between rice and maize at the moderate threshold. Next, the scGEN model was initialized using scgen.SCGEN, and trained using scgen.model.train, using default parameters. Next, the integration was performed using scgen.model.batch_removal. To evaluate the integration beyond the low dimensional representation, MetaNeighbor was used to compare the post integration similarity of cell types (Crow et al. 2018). In order to confirm the model was utilizing coexpression proxies and not relying on training information, the integration was run 872 times, starting with random gene pairs. Following each

run,7 random pairs were replaced with 7 coexpression proxies, until all were replaced. GO term enrichment was performed using Fisher's exact test from scipy.stats.fisher_exact() to find terms over-represented in the coexpression proxies, utilizing all genes in the bulk network as the background gene set. Multiple hypothesis correction was performed using the Benjamini-Hochberg correction function from statsmodels.stats.multitest.multipletests() at alpha .05.

Data Availability

All analyses were performed in Python 3.9, Pandas 1.5,SCGEN 2.1.0, Statsmodels 0.14.1, Scipy 1.12.0, and SCANPY version 1.9.1(Wolf, Angerer, and Theis 2018). Aggregate coexpression networks were downloaded from CoCoCoNet (J. Lee et al. 2020). *Arabidopsis thaliana* single-cell RNA-seq expression data totaling 16636 cells from 4 datasets were downloaded from the Gene Expression Omnibus GEO IDs: GSE116614, GSE121619, GSE123818, GSE123013)(Denyer et al. 2019; Ryu et al. 2019; Shulse et al. 2019; Jean-Baptiste et al. 2019). Cluster assignments were downloaded from GEO for IDs GSE121619 and GSE123013, or provided by the authors for IDs GSE123981 and GSE116614. *Oryza sativa* single-cell RNA-seq expression data and accompanying cluster assignments were downloaded from GSE146035(Liu et al. 2021). *Zea mays* single-cell RNA-seq expression data and accompanying cluster assignments were downloaded from GSE183171(Li et al. 2022), and only nitrate treated cells were used. Orthology information is from OrthoDBv11(https://www.orthodb.org/).

Chapter 3

Orthology level comparisons reveal conserved gene coexpression obscured by gene duplication

In the previous chapter, we identified a lack of one-to-one gene mappings between plant species as a key issue impeding the integration and comparison of cross-species plant data. In previous work, our group has found that conservation of coexpression between plant species tends to be lower and to drop more rapidly than conservation of coexpression in mammalian species (Crow et al. 2022). However, we note that in plant genetics redundancy is common, and plant species like Arabidopsis often require multiple knockouts to show a phenotype. Here, we hypothesize that plant gene coexpression is especially malleable between plant species, and that these slight shifts (shifting coexpression to a highly similar gene in the same family) may be driving the drop in conservation of coexpression. To address this, we build upon our group's previous work to propose an orthogroup-wise conservation of coexpression framework to better capture major shifts in coexpression without them being washed out by these minor shifts.

3.1 Introduction

Groups of genes are expressed simultaneously to accomplish complex biochemical functions and finely tuned developmental programs. From the earliest high throughput quantification of mRNA, analysis turned towards examining this coexpression to identify shared functions and regulation (Eisen et al. 1998). Coexpression based analysis has been critical for the annotation of gene function in model organisms, prediction of protein interaction, and discovery of target genes for drug development (Hughes et al. 2000; Drewes and Bouwmeester 2003; Noort, Snel, and Huynen 2003; Walker 2001). Previous work has shown that one-to-one homologs retain their coexpression partners between species (Crow et al. 2022). When multiples copies of duplicated genes are retained, coexpression patterns between the newly generated paralogs typically diverge (Crow et al. 2022). Often, one member of the pair rapidly shifts its coexpression pattern while the other member retains its ancestral pattern (Birchler and Yang 2022). While much research has focused on the ultimate fate of the diverging gene — its subfunctionalization, neofunctionalization, or loss — less attention has been given to the new set of genes the newly generated paralog expresses with. Given time, these diverging networks adapt to new environments, specializing and developing unique and useful phenotypes.

While all evolutionary changes are downstream from sequence changes, there are two main categories of changes that create the unique phenotypes that define different species: changes to the coding sequence of genes, and changes to

gene regulation. While sequence is mostly static during the life of an organism, gene regulation is constantly shifting to respond to environmental changes, prepare for future life stages, and minimize pathogenic disruptions. Because of this, it is difficult to quantify changes in regulation between species with any single assay, which is made doubly difficult by upstream changes in genetic sequence. Understanding and predicting how a group of genes will respond to perturbation is critical for transferring specialized knowledge and traits from one species to others in which they are desired. With the increase of generating both genetic and expression data for novel species, the community now has access to a rich tableau of useful traits. However, this broader sampling of genetic diversity brings two challenges. One, that even without a major phenotypic change, genetic networks can have compensatory changes that make direct comparisons more difficult. Two, that as the evolutionary distance between two species grows, so too does the chance that one of the species will have experienced a gene duplication, clouding the ability to make direct comparisons between the two species.

Natural selection occurs based on phenotype, not genotype. When multiple slightly diverged copies of a gene exist, it can rapidly become unclear how exactly each gene contributes to the phenotype (Pickett and Meeks-Wagner 1995). Even when each gene's contribution in a standard background is known, the exact response to a genetic perturbation can be impossible to predict. Depending on the history of each gene copy following duplication, one gene may compensate for the loss of the other, or one may be completely nonfunctional. Extending this scenario across species, the same trait in different species may have different degrees of

polygenicity, and different levels of redundancy. While recent work has shown that pleiotropy tends to be conserved across species, the same work has highlighted that cis-regulatory regions are not conserved across long distances (Hendelman et al. 2021). This lack of cis-regulatory conservation further reduces our ability to predict the degree of polygenicity and redundancy of a trait in a new organism based on model species. This complexity makes comparisons of function across species difficult – not only does it obscure which genes should be compared, but it also introduces scenarios where genes can show low conservation due to their function being split across multiple genes.

Despite these challenges, predicting changes in expression in a novel species based on a related species expression response remains critical to transferring useful traits, understanding responses to stress, and developing modifications to existing genomes that improve productivity. While advances in genomic technology like CRISPR have improved our ability to finely dissect the way changes in non-coding regulatory regions upstream and within genes affect gene regulation, this process is still slow (Xingang Wang et al. 2021). The characterization of genes requires the generation of multiple transgenic plants, a slow process that requires transfection and (typically) regeneration from tissue culture. Additionally, characterizing an individual gene's impacts does not provide information about the other genes in the family, the degree to which they might be able to buffer changes in expression in that gene, and the impacts of other genetic background mutations in the species. Additionally, the large shift in genetic background between species means that characterization of a single gene within a

species may not characterize a similar gene in another species. Instead of characterizing the relationship of an individual gene to each other gene, we would like to identify the underlying relationships between groups of genes that are deeply conserved, eschewing the more minor shifts between genes. However, identifying what is deeply conserved in plants without making only surface conclusions is difficult.

In an ideal world, we would be able to predict gene expression changes in response to perturbation in a new species based on the responses of the ortholog of that gene in a different species. However, this approach does not work in plants due to the rapid expansion of gene families and rapid adaptation and radiation of angiosperms (Y. Zhang et al. 2017). We suspected that in plants, frequent gene family expansions were enabling the fine tuning of gene modules by allowing shifts in expression between very similar genes. However, we noted that the underlying requirements for a successful plant (convert light into sugar, maintain homeostasis, resist pathogens) are not shifting and that plant evolution is highly driven by morphological innovation, which is often regulatory. Based on this, we hypothesized that plant gene modules were mostly determined by highly conserved sampling patterns from the same families. As a toy example, imagine that two plant species genes from gene families Alpha, Beta, and Gamma are coexpressed to accomplish a key task like photosynthesis. In a common ancestor, there might have been only one gene in each family. Then, a genome triplication event might have created 3 genes in each species while also triggering the species to diverge. Next, each species picks a gene from each family to express in leaves, and a different

gene from each family to express in stems. Because of this, genes from one species would show low conservation of coexpression between the species, despite the underlying coexpression patterns (Alpha, Beta, and Gamma families being coexpressed for photosynthesis) being deeply conserved. In reality, angiosperms have undergone many whole genome duplications, and which exact genes are sampled for the task and what tissues they express in is highly variable. This creates a scenario where many coexpression changes we capture between species are false positives and wash out signal from real shifts in the underlying module space. To address this scenario, we needed to find a way to capture the relationship between gene families, which we believe to be highly conserved, instead of capturing gene-gene relationships, which are more variable due to highly expanded gene families.

To improve our ability to predict how gene networks function across species, we developed a framework, orthogroup-wise conservation of coexpression, that leverages known gene relationships between species to smooth the background shifts in gene regulation between species. This framework builds on previous work in our group, where we developed conservation of coexpression to evaluate how functionally similar two genes are between species based on their coexpression partners (Crow et al. 2022). In our updated framework, we merge gene families, averaging their expression before generating coexpression networks using the relationships between averaged families. While several approaches were possible, averaging the gene expression across all genes in the families stood out to us as a simple, unsupervised approach that could scale to all our available data. The first alternative approach would be to annotate existing bulk RNA-seq data by

tissue and use it to generate tissue specific coexpression networks. Using tissue specific coexpression networks, we could attempt to finely dissect gene families, matching genes more accurately. However, this approach would require the hand annotation of thousands of samples. Another alternative approach would be to generate coexpression networks using single-cell data specific to one cell type. With this approach, we could directly compare genes in homologous cell types while also identifying which cell types expressed which genes. This would allow us to directly match genes without hand annotation. However, single cell data is available for only a handful of plant species, and it is often limited to root tissue. Because of these limitations, we decided to stick with the simple approach of averaging all genes within a family to get an average family expression in each sample, and then generating the coexpression networks from these averaged samples. This approach enables us to better capture real shifts in the network, as opposed to minor finetuning and rewiring.

While network rewiring is common as species diverge, we hypothesized that this rewiring mostly occurs with genes closely related to the ancestral gene's coexpression partners. That is, relationships between gene families tend to be conserved, even if the specific relationships shift. We believe that this is because it is rare for the underlying network of gene family relationships to shift, and that most shifts in coexpression between plant species are occurring within these families. Thus, by smoothing these within families shifts by averaging family expression, we can improve the amount of signal we capture in conservation of coexpression by instead measuring the conservation of coexpression between

families. Indeed, in recent work in vertebrae has also shown that expression profiles are more conserved at the orthogroup level than the individual level (Marlétaz et al. 2018). We show that coexpression networks built using gene families are highly robust across large evolutionary distances, allowing us to improve the signal for identifying conservation at large evolutionary distances. This highlights our framework's robustness, and its value at both predicting responses to perturbation, and at identifying true, consequential changes in the underlying regulatory network that might otherwise be hidden by the many inconsequential shifts between gene family members.

3.2 Results

Angiosperms have a nearly infinite range of desirable traits and extremely frequent gene duplications, with both whole genome duplications and tandem gene duplications occurring in many lineages. This makes angiosperm lineages an ideal venue to evaluate our framework for better capturing the underlying coexpression network resulting from conserved genetic modules. Because of the frequent gene rewiring in species, attempting to predict shifts in gene expression based on shifts in a related species does not work except in the closest related species. Often, it fails even between accessions.

In a toy example, we highlight the logic behind how our framework reveals the underlying shared coexpression network, simplifying expanded families and highlighting both where the network has not meaningfully diverged and where a critical change has occurred (Fig 3A). Briefly, instead of a traditional calculation

of coexpression by calculating pairwise Spearman correlation coefficients for each gene pair, we instead average all genes within an orthogroup, generating a coexpression matrix between the averaged values. In the toy example, each shape represents a gene family. For gene families with a duplication, the two different genes in the family are represented with colors. In Species A, the circle and diamond genes have been duplicated, and in species B, the square and triangle genes have been duplicated. Edges represent coexpression relationships, with thickness representing the strength of the relationship. In both species, the exact relationships and the strength of those relationships varies in the gene coexpression network. However, both species have two submodules in them. Using a traditional coexpression approach, it appears that both coexpression submodules have diverged between the two species, as the coexpression relationships have broken up due to gene duplication causing subfunctionalization. To mitigate the impacts of this, we apply our framework, generating the bottom network. This approach merges genes within gene families, resulting in the purple shapes. The orange shapes, which have only one gene per family, are not merged. The resulting network makes it clear that while the top submodule is actually conserved between the two species, the bottom module has diverged. Highlighted in red, the coexpression between the cross and the diamond in species A has stopped in species B, and the cross is instead coexpressed with the triangle family. This highlights how our approach can improve signal, highlighting true shifts in the coexpression network by reducing noise from minor rewiring within families. While simple, this approach rests on the observation that the underlying

coexpression network that drives most processes is highly conserved, with most cell types sampling from this network.

To underscore the choice of angiosperms as a model for evaluating this framework, we compare the number of one-to-one gene pairs available for making cross-species comparisons between *Homo sapiens* and related mammalian lineages to those available between *Arabidopsis thaliana* and related lineages (Fig 3B). To do this, we select 12 mammals and 7 angiosperms (all species for which we have a comparable amount of data to construct high quality coexpression networks) at a range of distances from our two starting species. These groups are similar in age, with mammals first appearing 178 million years ago, and angiosperms appearing 135 million years ago. For each selected species, we retrieve all gene orthology relationships to the starting species for OrthoDB V11(Kuznetsov et al. 2023). Using these relationships, we identify one-to-one gene pairs by removing all pairs with multiple matches between species, leaving only one-to-one genes. We plot the total number of one-to-ones available from the root species, highlighting the disparity between the two clades. Even though there are more Arabidopsis genes than human genes, we plot these as total pair counts, instead of as a percentage of the genome. In addition to highlighting the dearth of one-to-one genes in Arabidopsis, because genome expansions would rapidly drop the percentage of genome aligned for larger species, this metric is fairer. The massive difference in one-to-one gene pairs in Arabidopsis relatives compared to human relatives shows why angiosperms are the ideal lineage for testing our framework. Unlike in mammals, plants rapidly expand their gene families, so they

serve as an excellent model for a method to improve comparisons by reducing noise from rewiring following gene duplications. In comparison, most mammal gene families are relatively small, with few highly similar or redundant genes. As a result, averaging within mammalian families is unlikely to dramatically improve cross-species comparisons, in contrast to plants, where we can merge highly similar genes based on family.

Next, we highlight a major issue the field faces – that while one-to-one gene pairs shows high conservation in the coexpression space, this approach to capturing similarity rapidly drops in genes that are N-to-M (having expanded gene families) between two species. Currently, this is taken to indicate that these genes have had a large shift in function (Crow et al. 2022). Yet, we know that many of these genes are recent duplicates and may be very similar in function. We evaluate gene function similarity across species using a "conservation of coexpression" approach. This method compares genes by assessing the similarity of their coexpression partners. First, we identify the top 10 coexpression partners for each gene in Species A, focusing only on genes with one-to-one orthologs in Species B. This set of partners serves as a "functional fingerprint" for the gene, based on guilt-by-association principles. We then use these fingerprints from Species A to predict the most coexpressed genes for each gene in Species B. This prediction generates an AUROC score for every gene in Species B, representing its functional similarity to the corresponding gene in Species A. Next, the entire process is reversed, using Species B data to predict coexpression in Species A. Finally, we average the scores from both directions to create a final coexpression conservation score. This results

in an AUROC, where .5 is random, 1 is perfectly conserved (same top coexpressed 10 partners) and 0 is completely flipped (top 10 partners are bottom 10 in the second species). We perform this between Arabidopsis and the previously selected species. Using this metric, we show that as expected, one-to-one genes are conserved between species (Fig 3C). In contrast, N-M genes rapidly lose conservation between species. One-to-one genes have an average conservation of coexpression of nearly .9, while N-to-M genes have an average conservation of approximately .78. This dramatic difference highlights how rapidly the genewise conservation of coexpression metric drops for N-to-M genes. While we expect one-to-one genes to be more conserved, as these genes are from families where gene duplications were rapidly lost, we believe that many N-to-M genes should also show high conservation, as they often retain highly similar functions. Additionally, we know that in plants duplicated genes can frequently act as backups for other similar genes, requiring multiplexed knockouts to achieve a phenotype. Because of this, we think that this degradation in our conservation metric is mostly a function of rewiring between genes within families, and by mitigating this, we can better capture significant shifts in the coexpression network.

Before moving forward with evaluating an improved framework, we need to confirm that conservation of coexpression is accurately capturing real shifts in gene function between species. To do this, we investigate conservation of coexpression as a function of species divergence time from Arabidopsis. If conservation of coexpression is accurately capturing divergence between species, conservation of coexpression should decline with species distance. Taking

divergence distance from Time Tree (an online database of species divergence times), we calculate conservation of coexpression for all one-to-one genes between Arabidopsis and other plant species (Kumar et al. 2022). We show that as the time since divergence increases, the degree of conservation of coexpression drops, confirming our method's ability to capture true divergence in function, which we know increases as species become more distance (Fig 3D). While the drop is subtle, it is important to remember that these genes are those that are critical enough that any gene duplications have been purged from the genome. As a result, they are highly conserved, so any shifts in conservation of coexpression are likely to be minor. Indeed, even across 100 million years of divergence time, there is only a mean difference of .03. However, the distribution of the groups does shift more dramatically as divergence time increases.

Next, we evaluate our orthogroup-wise conservation of coexpression framework, comparing the existing genewise conservation of coexpression calculation to our new method, where we merge genes based on gene family information. To order individual orthogroups by age, we sort them by the number of species which have genes in that orthogroup. Older orthogroups, which represent gene families established earlier, have more species with genes in that orthogroup. This is because gene families are shared by all species that diverged after the creation of the gene family. For each orthogroup, we calculate genewise conservation of coexpression between all cross-species from Arabidopsis to other species for gene pairs in the orthogroup. We then calculate the orthogroup wise conservation of coexpression for the orthogroup between the same sets of species.

This is done for all orthogroups, and the results are plotted by the number of species in the orthogroup, a proxy for age. We show that genewise conservation of coexpression rapidly drops as orthogroup age declines, as expected. This decline is rapid, and the oldest orthogroups have a mean conservation of coexpression score of about .7, which is low for a relatively recent group like angiosperms (which first appeared ~140 MYA). In contrast, we show that when we use our new framework for calculating orthogroup-wise conservation of coexpression, the scores decline much more slowly (Fig 4). Using the orthogroup-wise conservation of coexpression, even the oldest orthogroups maintain a conservation of coexpression score of approximately 0.85. This is a stark contrast to the genewise conservation of coexpression, which bottoms out around 0.75. Compared to the total loss difference in one-to-one genes across 160 million years of evolution, 0.03, the difference between orthogroup-wise conservation of coexpression and gene-wise is massive. This matches expectations, as while genes are diverging over time, much of the divergence captured by genewise conservation of coexpression is *actually* rewiring between genes in the same families and does not represent major shifts in coexpression. By instead capturing shifts in coexpression only when they change relationships between entire gene families, we highlight that much of the loss of conservation of coexpression is instead rewiring between families, and that by averaging these families we can better capture the true underlying coexpression relationships between genes.

Next, we extend our method to investigate if merging other groups of genes provides a useful framework for investigating the evolution of traits. Instead of

clustering by orthology, which puts genes with highly similar genes together, we investigate if we can instead cluster by Gene Ontology (GO) terms (Aleksander et al. 2023). While a similar approach has been taken to compare cellular profiles between species, it has not focused on identifying functions that are rapidly evolving in a broad category of species, like angiosperms (Song et al. 2024). Like orthology groups, GO terms are comparable across species and cluster genes with similar functions. Unlike orthology groups, a gene can be annotated with multiple GO terms. Additionally, GO terms do not cluster by just gene origin, but also by "function" – by what the gene "accomplishes". For example, GO terms exist for important organismal functions, like photosynthesis. These terms can link together genes with very different evolutionary origins that are coexpressed to achieve a role. Ideally, we would focus on only these GO terms, as they cover a different space than the orthology based grouping. To sort these useful GO terms from other, less interesting GO terms, we evaluate all GO terms based on their intergroup coexpression, using EGAD (Ballouz et al. 2017). Briefly, EGAD evaluates whether a set of genes is more coexpressed with each other than expected by chance based on how many genes they are typically coexpressed with. Thus, it gives a good metric for the modularity of a given list of genes.

Evaluating all GO terms across 13 plant species, we identify highly modular GO terms. For all of these GO terms, we then calculate the average gene-wise conservation of coexpression between all genes annotated in the group, averaging their conservation of coexpression. Highly conserved genes should have high conservation of coexpression in all species. Next, to check that these groups

cover the same space of genes generated by and independent coexpression based clustering of genes, we use WCGNA to cluster Arabidopsis genes (B. Zhang and Horvath 2005). Using these groups, we perform the same evaluation of the genes as the GO term evaluation and plot them with our GO terms (Fig 6A). We show that our GO terms clusters cover a similar range of modularity and conservation as genes clustered by coexpression space, confirming that we can use GO terms to capture the same types of clusters we capture with unsupervised clustering. Here, we use modularity to determine which clusters we are interested in. In this space, we are interested in highly modular clusters, as those GO terms are those that are actually coexpressed with each other, as opposed to GO terms linked together for other reasons. For example, a GO term with low modularity but high coexpression might be DNA repair genes, which are expressed at different times but share a biochemical function. Among highly modular genes, we are most interested in those that are either highly conserved or those that have very low conservation. While our interest in highly conserved genes is obvious, genes that are poorly conserved indicate they are an area of rapid innovation across plant species. These genes are critical to identify, as they are genes that define differences between plant species.

To meaningfully evaluate what is occurring in our gene coexpression space, we hand classify over 3000 GO groups. We put all GO terms into one of 6 categories: photosynthesis, ribosomal, hormonal, secondary metabolism, primary metabolism, or other. In plant data, ribosomal and photosynthesis genes are known to be highly conserved, and our approach of clustering genes by GO terms confirms

this. These genes end up on the top right of the graph (Fig 6B). On the bottom right, where strong modules that are experiencing rapid shifts in coexpression partners are located, we find hormonal and secondary metabolism genes. Hormonal genes are important for defining plant architecture, a highly flexible trait that varies strongly within genus and even within species. Secondary metabolism genes are also known to be an area of rapid innovation, as they often are important for responding to pest stress and abiotic stress. Successfully identifying these groups of genes as highly conserved and rapidly shifting highlights this approach as a useful way to rapidly spot where innovation is occurring between plant species.

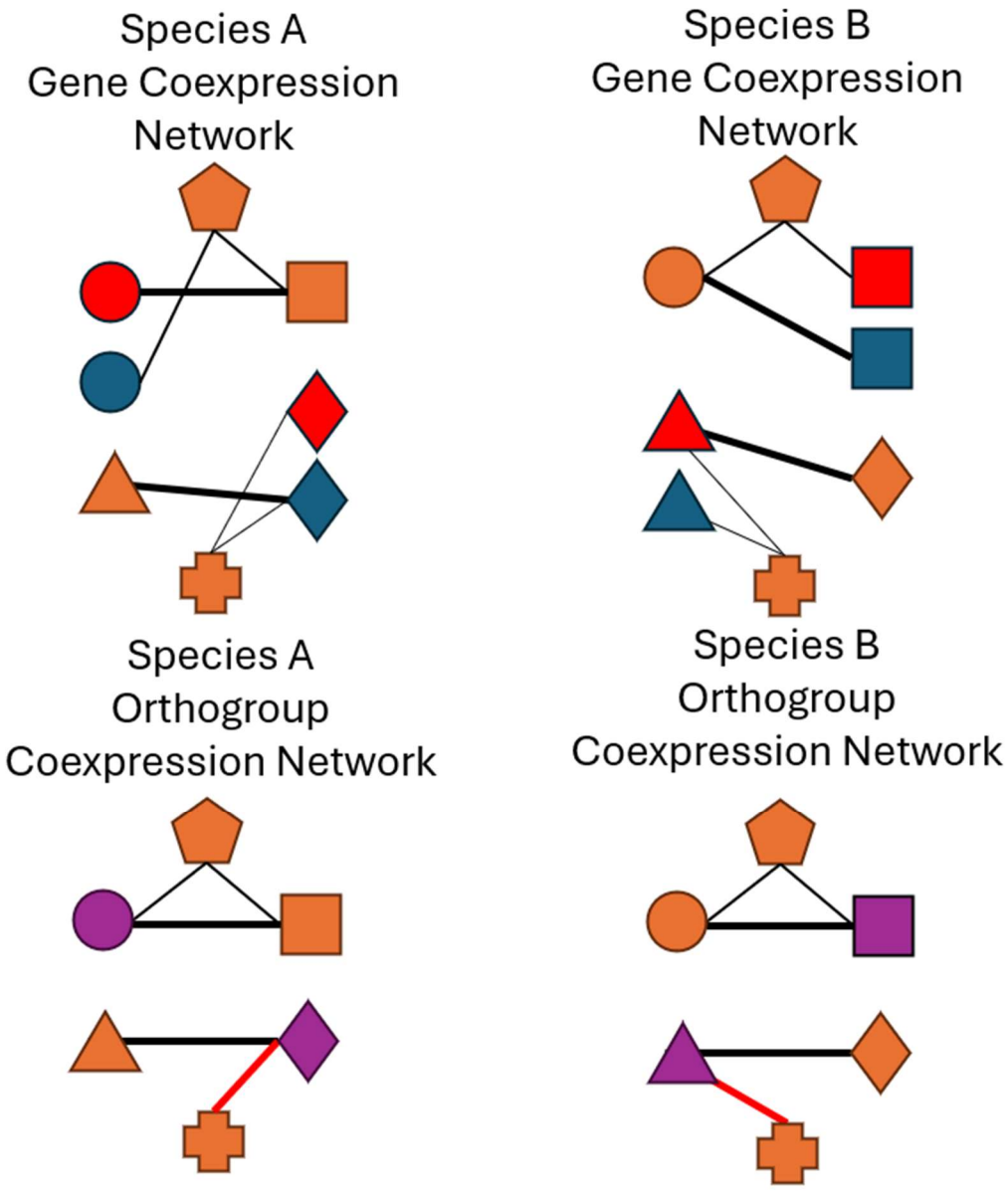


Figure 3A. *Schematic depicting how merging gene coexpression networks into orthogroup-wise coexpression networks can improve the signal for detecting true changes to the underlying coexpression patterns between gene families. Red shows true coexpression changes to underlying family relationships*

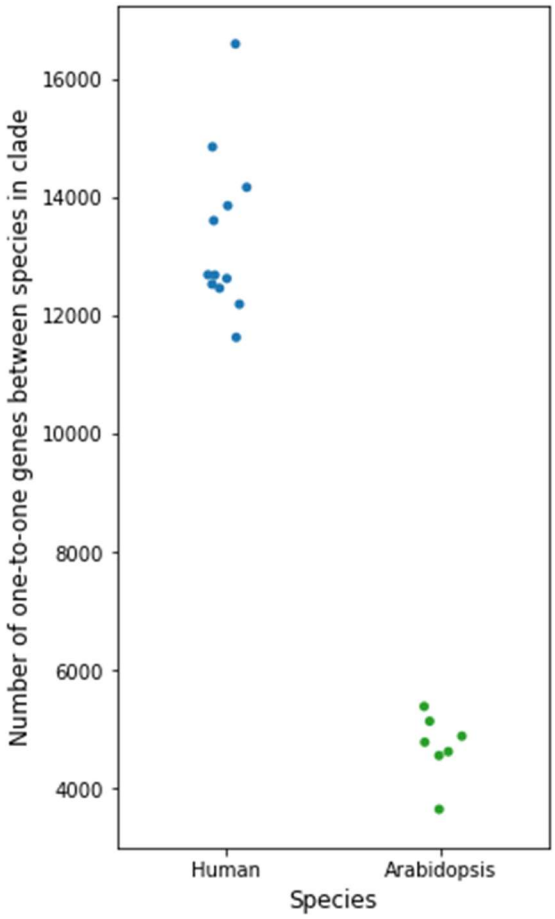


Figure 3B. Strip plot comparing the number of one-to-one gene pairs between humans and other mammals, and between Arabidopsis and other angiosperms.

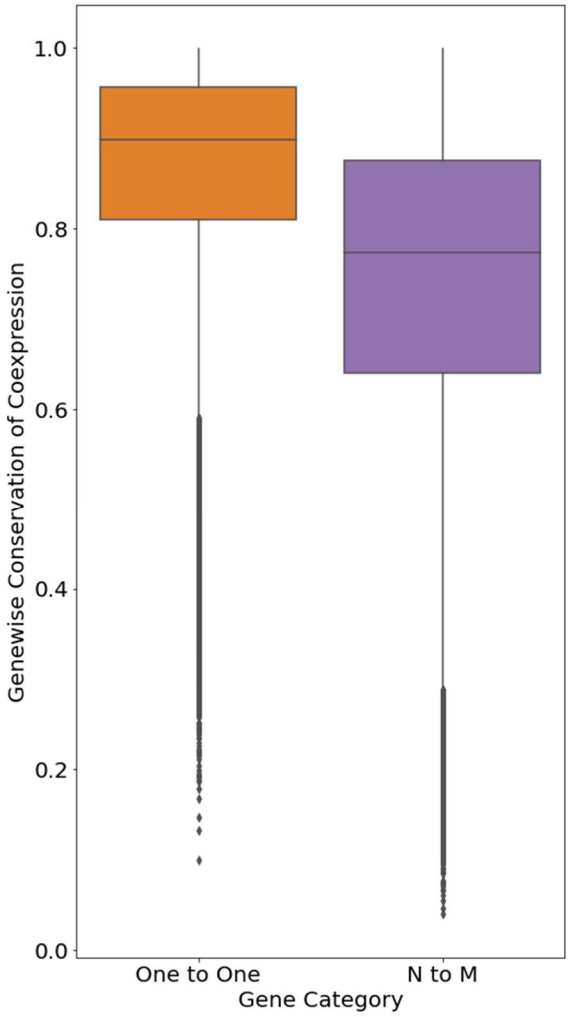


Figure 3C. Comparison of conservation of coexpression scores of Arabidopsis genes to all other orthologous genes in selected species, separated by if the gene has only one ortholog in the selected species and in Arabidopsis, or if it has multiple orthologs.

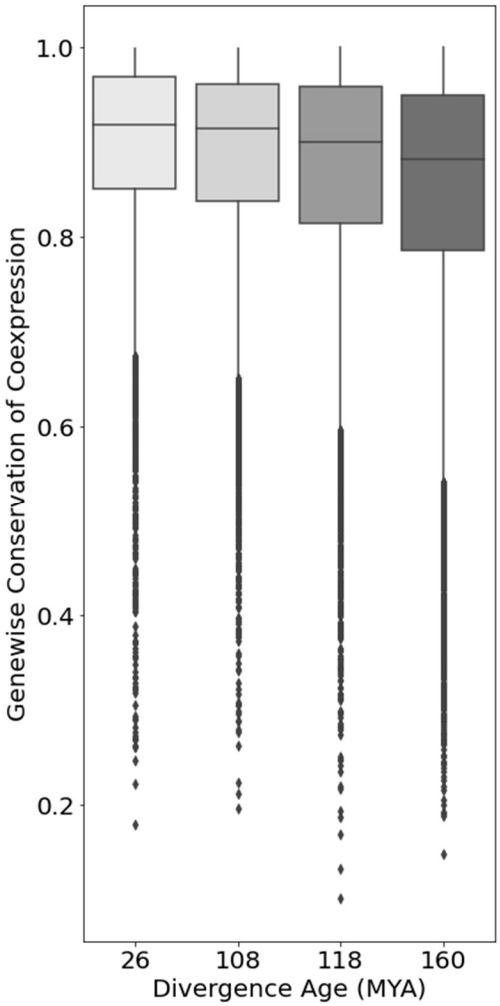


Figure 3D. Boxplot showing different conservation of coexpression scores for all genes as species get more distant from Arabidopsis. Distinct groups are from clustering of species at specific distances (e.g. All monocots are at 160 MYA).

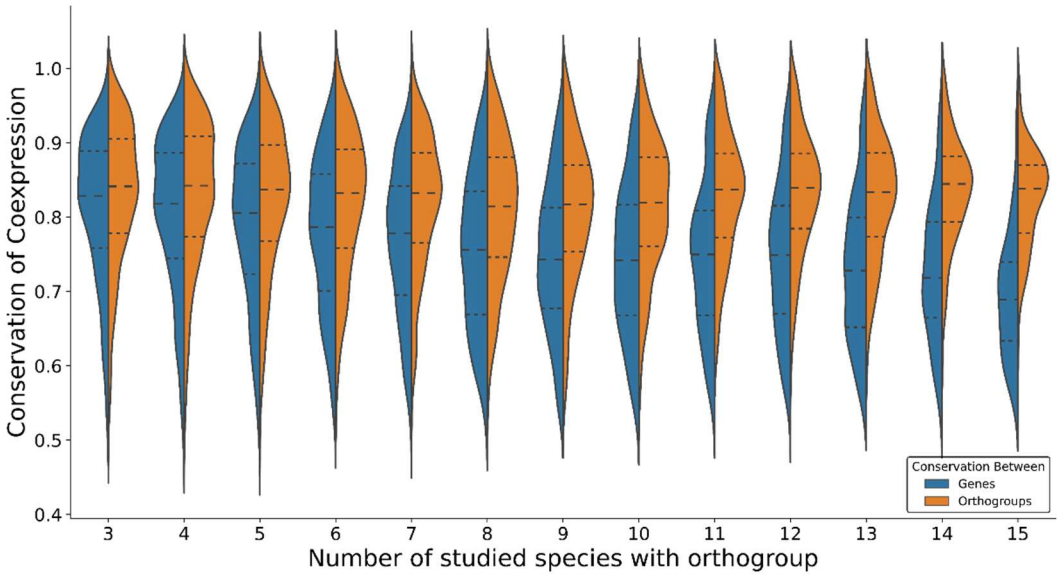


Figure 4. Violin plots of genewise and orthogroup-wise conservation of coexpression versus Arabidopsis genes for orthogroups with different numbers of species present in each orthogroup (a proxy for age).

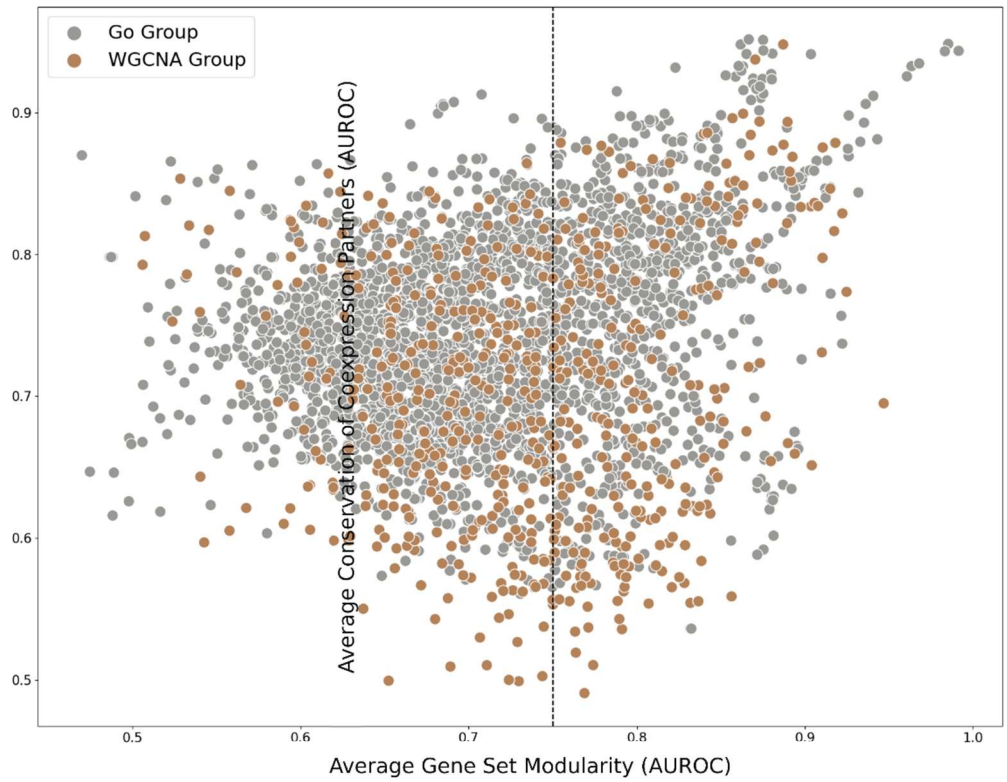


Figure 6A. Plot of gene set modularity vs genewise conservation of coexpression for both GO terms and WCGNA defined groups of genes.

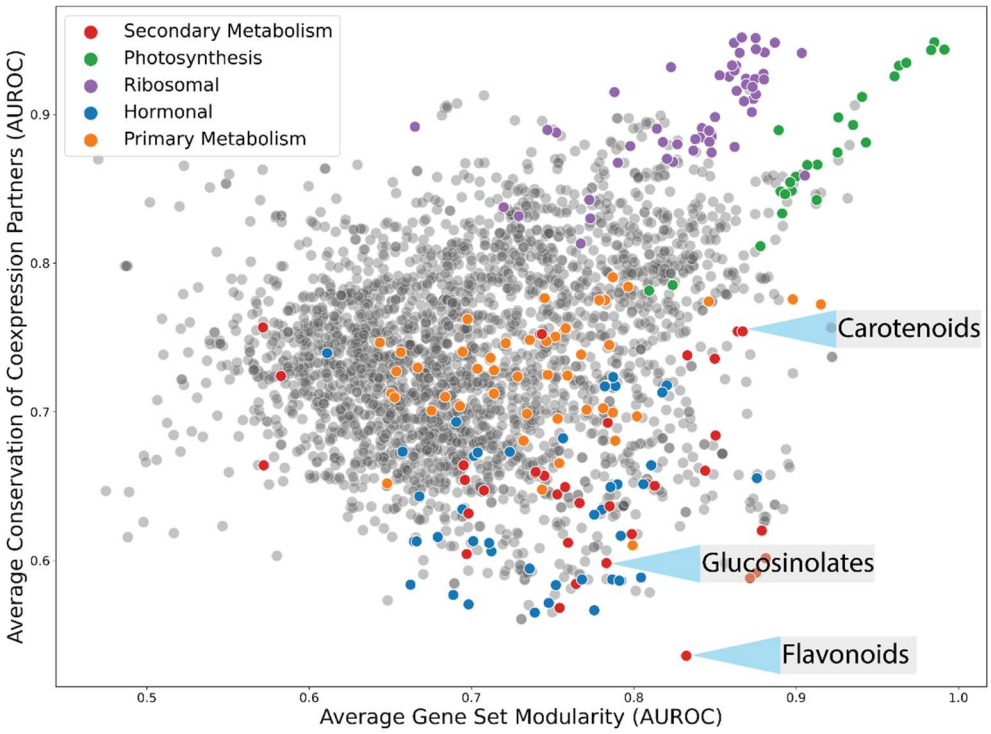


Figure 6B. Hand annotation of GO terms identifies rapidly rewired GO terms as well as deeply conserved terms.

3.3 Discussion

Our work introduces a novel framework for analyzing coexpression relationships between genes. By merging genes based on their orthology relationships we can minimize shifts in coexpression relationships within families, revealing more substantial shifts that occur between families of genes. This approach is especially relevant in plants, where frequent whole genome duplications and rapid radiation have resulted in large gene families that show rapid loss of conservation of coexpression. When we merge these families, we reveal that much of the captured change in conservation is occurring within gene families, as gene expression is finely tuned to better match the environment, abiotic stress, and pests facing the species. While previous work has noted that duplicated genes tend to diverge rapidly, it had not been investigated what types of shifts in expression pattern were underlying this divergence (Crow et al. 2022). Our work identifies many of these shifts as minor, within-family changes. By averaging away these minor within-family shifts, we leave only shifts in coexpression outside of families, which indicate much more substantial changes to the underlying coexpression network.

We further leverage this simple, scalable approach to cross-species coexpression comparisons by applying it to GO terms. Unlike purely family based orthology databases such as OrthoDB, GO annotations include cellular, biological process, and molecular function annotations. These groups cluster genes not just by their origin, but also by what they do both biochemically and within the context

of the cell. By merging genes within these clusters and evaluating their conservation of coexpression between species, we can rapidly reveal what gene functions are highly conserved between species and which functions are areas of rapid innovation. These clusters, as well as their human readable annotations, enable us to quickly categorize gene groups into broader categories to make broad conclusions about where innovation is occurring between species. By hand annotating GO groups, a quick and easy process in comparison to many other types of hand annotation, we highlight hormonal and secondary metabolism genes as areas of rapid innovation in plants. This result is backed up by our method also picking up the deep conservation of ribosomal genes and photosynthesis genes. Photosynthesis is the defining feature of plants, and ribosomal genes are a deeply conserved set of genes that are shared by all living organisms. In addition to its application to GO terms, this approach can be used for any gene set, such as a set of genes identified from a screen, in order to identify how rapidly the set is diverging between species.

These findings have important implications in the field of plant genetics. While sequencing costs for investigating perturbations in crop species continue to drop, other issues, like extraction protocols, uneven germination, and complex, polyploid genomes with mulitmappers can make using RNA-seq in a crop species of interest more difficult. By improving our ability to capture conservation between species and differentiate highly conserved gene modules from less conserved ones, we open the door to easier prototyping of modifications to plant species via model organisms before deployment to other related species.

A second critical implication of this study is what it says about how plant genes evolve and acquire new functions. We identified that much of the coexpression divergence in plant genes results from shifting coexpression to a highly similar gene, fine tuning the network but not making major changes. In the context of plant evolution, this makes sense. Much of the rapid radiation of angiosperms involved adaption a successful model to new habitats (Ren et al. 2018). This adaptation did not require massive shifts to what the organism fundamentally does to survive. Instead, it often required architectural changes that allow the plant to better tolerate the abiotic stresses present in the environment – things such as wind, sun, or drought. By identifying this source of coexpression conservation noise, we are better able to model how plants evolve to fit new environments and understand Darwin's abominable mystery.

While this work has laid out a new framework for analyzing plant coexpression data, there are several key limitations to keep in mind. This framework is unlikely to be anywhere near as powerful in organisms without frequent gene duplication, such as mammals. In these species, the infrequency of gene duplication does not leave much room for fine retuning between similar genes, and as such merging genes my family is more likely to instead obscure real shifts in the coexpression network. In species like these, the abundant one-to-one gene pairs are ample for prediction gene response across species. Additionally, it is unclear whether the lack of conservation for secondary metabolites is an accurate conclusion to draw from our GO term analysis. Secondary metabolites are often produced from completely different pathways in different species, so low

conservation may be from comparing different genes across species (Bennett and Wallsgrove 1994). This is because unlike abiotic stressors, biotic stress sources like pests are in an evolutionary arms race with plants, and as such many plant families develop idiosyncratic ways to combat them. For example, glucosinolates are an important chemical group in Brassicaceae that improve pest tolerance (Mitreiter and Gigolashvili 2021). Because this response originated in Brassicaceae, it is unlikely that our method would be able to accurately capture conservation in this family, as these species are likely instead adapted to use this new chemical response.

Chapter 4

Leveraging coexpression in plants for comparative plant biology approaches

In the previous two chapters, we proposed and evaluated two new frameworks for improving cross-species comparisons in plants using coexpression. Here, we will walk through three real world applications of coexpression analysis for cross-species comparisons in plants, one of which employs EPIPHITES to improve integration (Chapter 4.3). These vignettes, all drawn from work we published with our collaborators, highlight the power of coexpression for answering complex questions in plant biology.

4.1 Introduction

In the last decade, comparative approaches to plant biology have become increasingly important to understanding the complex interactions between genes and how their products impact plant development, metabolism, and the response to environmental stimuli. While traditional comparative analysis between matched genes has enabled key discoveries in plants, the dearth of one-to-one genes makes this type of analysis difficult at larger evolutionary distances or with gene families that have undergone repeated expansion. Coexpression based analysis has been a critical tool in improving cross-species comparisons in plants, offering insights into functional shifts in genes between species as well as shifts in relationships between genes within a species. Coexpression analysis's power comes from its ability to detect changes not just in the molecular function of a gene, but also from capturing shifts in gene regulation between species, an important area of innovation critical to rapid adaptation to new habitats. As angiosperms rapidly radiated and colonized all 7 continents, modifications to architecture and development that enabled survival in new habitats were essential.

Understanding what changes occur and how they impact a plant's phenotype during evolution are two questions that comparative coexpression analysis is uniquely positioned to answer. Throughout my work, I have leveraged coexpression analysis to address a diverse range of questions in plant biology that leaned on cross-species analysis to improve the integration of cross-species data

for mapping rare cell types, to identify repurposed gene modules between species, and to identify key paralogs that are important to plant structural traits. Through collaborations bridging computational approaches and the collection of key data, we answered important questions about all of these areas, leveraging coexpression, scRNA-seq, and whole genome sequencing of novel model systems.

In the following sections, I will provide a detailed account of how I leveraged comparative coexpression analysis to enable these collaborative efforts. In these collaborations, we faced a range of challenges, including issues with normalization, integration, batch effects, and how to handle cross-species data where no exact homologs exist. Despite these challenges, the synergies of collaborative work allowed us to answer key questions in plant biology. Throughout this chapter, I will demonstrate how new analysis approaches, combined with technological advances such as scRNA-seq and cheaper whole genome sequencing, can push the field forward by enhancing the power and scope of analysis.

4.2 A pan-grass transcriptome reveals patterns of cellular divergence in crops

Cereal crops represent over 50 percent of calories consumed by humans, and are the foundation of the food supply for the entire world (Awika 2011). Across the globe, multiple cereal crops were domesticated in parallel, and now grains such as rice, wheat, maize, sorghum, and barley are key crops that cover millions of acres. These crops have a range of tolerances for various abiotic stressors, and key traits that differentiate them are often mediated by specialized cell types. One key trait, drought tolerance, is especially important in a warming world where access to water for irrigation is shifting. Sorghum is substantially drought tolerant, and one of the main reasons it is grown widely across the United States, Africa, and China is its ability to produce ample biomass and yield despite low water conditions (Hadebe, Modi, and Mabhaudhi 2017). In contrast, its close relative maize is much less drought tolerant, but produces a more palatable and widely useful grain.

These closely related crops with divergent traits presented a critical opportunity to employ comparative analysis to deepen our understanding of cellular evolution in plants. However, this analysis is complicated by a recent whole genome duplication in the maize lineage, likely as a result of allopolyploidizaiton (Yang et al. 2023). This whole genome duplication occurred several million years ago, but maize is still in the process of rediploidization, and

retains two sub-genomes. Additionally, maize transposons are unusually active, further complicating the shared gene space between the two species (Lisch 2012).

To help resolve this gene space, and to provide an outgroup that is likely more similar to the ancestral state of both grasses, we decided to include *Setaria viridis*, a related wild grass species. Seeking to investigate differences between the species at the cellular level, we generated two types of high-resolution RNA sequencing data, single cell data and single nucleus data. While single cell data tends to be higher depth, producing it is challenging, as it requires the breakdown of plant cell walls to generate delicate protoplasts prior to library generation. In contrast, single nuclei are much more robust and easier to generate than single cells. However, this improved robustness comes at a cost, as single nuclei contain fewer RNA transcripts, and may over profile some genes. To reap the benefits of both types of profiling, we generated both single cell and single nuclei data for roots in all 3 species, with single cell data providing depth of sequencing and single nuclei providing breadth, greatly increasing the number of cells profiled (Guillotin et al. 2023).

While generating two types of data provides the benefits of both types of data, it does introduce further issues. Combined with the initial design of our study, which involves data from 3 different species, we ended up with 6 unique datasets. With nuclei and cell based data from each species, we needed to appropriately integrate this data, handling the already challenging species specific effects in addition to potential differential responses to protoplasting protocols. In addition

to this complicated integration, comparing maize data to the other two species is complicated by its recent whole genome duplication, making it unclear which gene (or both) to compare to the single gene in either of its two relatives.

Addressing these issues was complicated, and took a multistage strategy. First, a key step performed by our collaborators was a substantial refinement of existing knowledge of maize gene homology. Although prior studies had linked homologs within the maize whole genome duplication, they had not further linked these genes to outgroup species. To enable cross-species comparisons, our colleagues aligned and identified these relationships. By making BLAST comparisons across both sub-genomes in maize and to both sorghum and *Setaria*, our colleagues identified one (sorghum) to one (*Setaria*) to one (maize sub-genome 1) to one (maize sub-genome 2) genes. Mapping these genes to the two sub-genomes was critical, and it enabled further downstream approaches we used for integration and batch correction.

Before we could perform any analysis, we needed to integrate and batch correct the 6 datasets, which spanned both species and data generation techniques. To do this in a robust way, we also needed to take multiple approaches in parallel, to ensure that any integration strategy we used was not driving spurious results. We tested a broad range of integration algorithms, and ultimately made several observations about integration of cross-species data with both single nuclei and single cell data. The key observation was that integrating these datasets across modalities required strict quality control of the single nuclei data. By its nature,

single nuclei data is lower depth than single cell data, and without careful selection of thresholds, this effect dominates any attempt at integration. To mitigate this effect, we needed to be stringent with quality control cutoffs for the number of genes detected per droplet. While using the same thresholds for nuclei and cells, we saw any attempt at integration completely dominated by this effect (Fig 6). In contrast, when we raised the quality thresholds for nuclei to require 3000 genes per cell, we immediately saw improved integration (Fig 7). Although this ended up removing about a third of the nuclei collected, the vast improvement in integration meant this trade off was worth it.

As part of the integration and batch correction process, a key step was selection of an integration algorithm. Most integration algorithms are designed to integrate data that are biological replicates with technological variance, from being done in different locations or during different runs. However, our integration scenario does not match this design parameter. By attempting to integrate across species, we confound technical variance with true biological variance. To unwind these, we need to make certain assumptions about our data and use integration algorithms that match these assumptions. We assume that technical factors impact all cell types equally, so that any technical variance will occur in all cells in the dataset. In contrast, biological variance should be mostly cell type specific, with different cells changing in different ways between the two species. Because of these assumptions, we need an integration algorithm that allows different integrations by cell type/cluster, instead of one that makes universal adjustments to all cell types. These types of integration methods tend to be “more aggressive”

– that is, they are more likely to fit away real biological differences while minimizing technical variance. Because of this, it was critical that we use multiple integration methods, in order to confirm that any results were robust to different integration approaches. Ultimately, we settled on using Seurat FindIntegrationAnchors function and scGEN's supervised integration in parallel to confirm that any results were robust to integration.

Having minimized technical variance between species and data collection methods, we next sought to identify where innovation was occurring within our two crop species. We used MetaNeighbor, which compares the transcriptional similarity of clusters of cells/nuclei based on highly variable genes, to compare all cell types in sorghum and maize back to *Setaria.* This allowed us to quantify how diverged each cell type was from the outgroup species, *Setaria*. While having only 3 species does not allow us to definitively state that any divergence is a result of evolution in one of the crop species, the rapid and major shifts occurring during artificial selection make this likely. Using this approach, we identified cortex and columella were two key areas of transcriptional divergence between the crop species and *Setaria* (Fig 8). To further investigate what was driving this divergence, we leveraged conservation of coexpression in existing bulk RNA-seq networks. We calculated conservation of coexpression between all genes expressed in columella and cortex in maize and sorghum, identifying 443 genes that were expressed in these tissues and showed high divergence in their conservation of coexpression, indicating that there was likely a functional shift occurring. Investigating these genes using GO term analysis, we found that they were highly

enriched for genes involve in the synthesis of key oligosaccharides used during mucilage synthesis. Root mucilage is an important mediator of plant interactions with the environment by shaping the microbiome and lubricating/controlling the hydration in the root soil interface (Deynze et al. 2018; Galloway, Knox, and Krause 2020; Kozlova et al. 2020). Examining all mucilage genes, we found that there had been a major shift in the location of mucilage synthesis in maize. We found that both *Setaria* and sorghum mostly synthesized mucilage in the cortex, while maize mucilage genes were expressed primarily in columella (Fig 9). This suggests that maize underwent a rapid cellular divergence in columella by recruiting a mucilage gene expression module from the ancestral expression location of the cortex. These genes were previously identified as being under selection during domestication, suggesting they had an important role in agriculture traits.

Using a comparative approach and leveraging both the wealth of existing plant data as well as novel data collection methods, we show that single-cell techniques can rapidly provide insights into plant cell-type evolution. We generated cell type specific data, and used it to rapidly identify where cellular divergence was occurring, before using coexpression analysis in the broader world of bulk RNA-seq based networks to generate a hypothesis for what might be driving this divergence. Returning to our single cell and nuclei data, we confirm this hypothesis by examining gene expression patterns in our species, confirming that mucilage genes were an area of rapid innovation in maize. This lays a path for

how future work to investigate how traits of interest have evolved in certain plant lineages can occur, enabling hypothesis generation and testing.

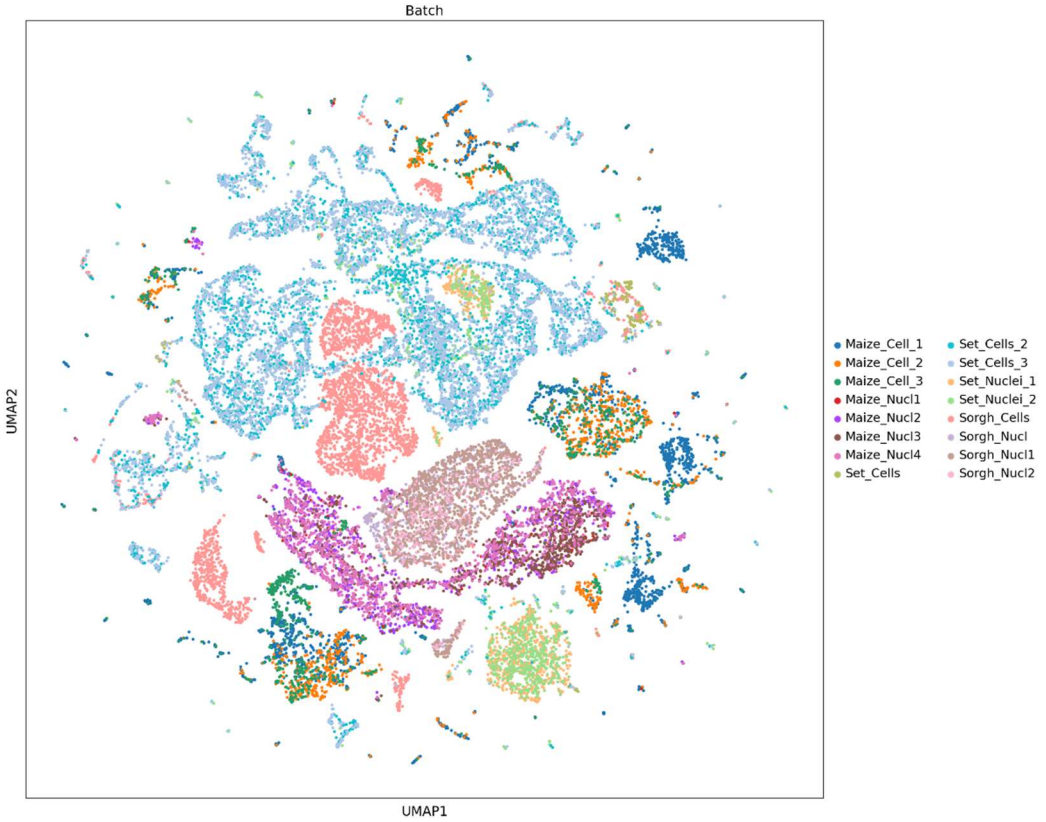


Figure 6. UMAP showing attempted integration of cross-species nuclei and cell data without removing low quality nuclei.

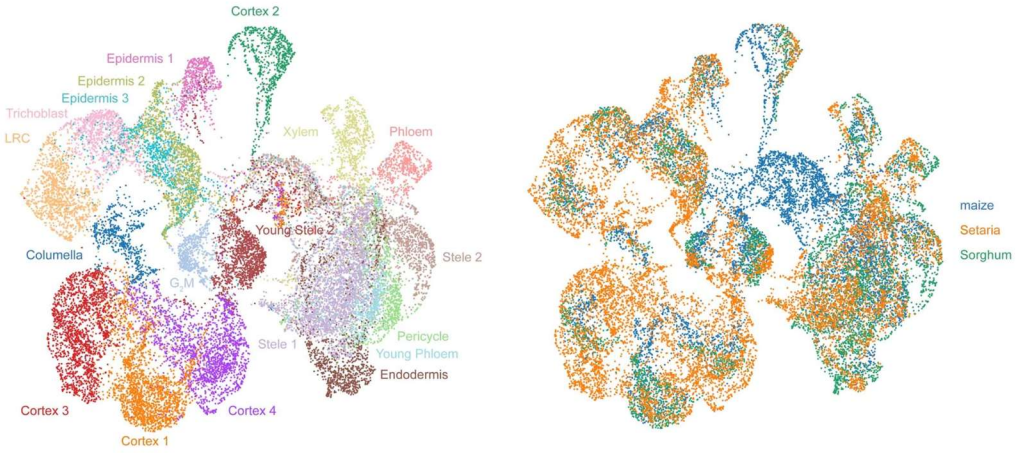


Figure 7. UMAP showing successful integration of cross-species cell and nuclei data following removal of low quality nuclei

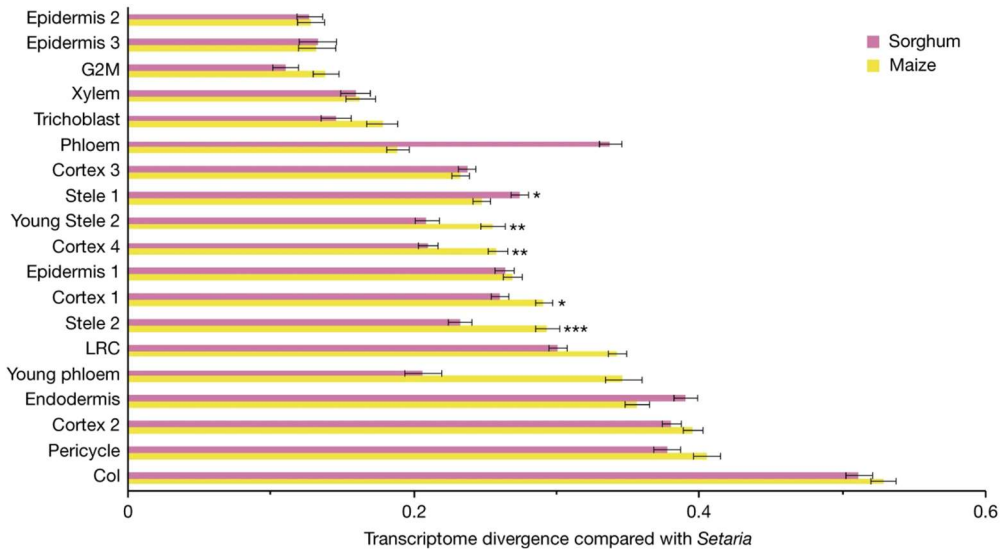


Figure 8. MetaNeighbor analysis showing a quantification of transcriptome divergence among cell types in maize and sorghum compared with the outgroup Setaria. *Two-sided Hanley McNeil test. *P < 0.05, **P < 0.01, ***P < 0.001. Error bars indicate standard error.*

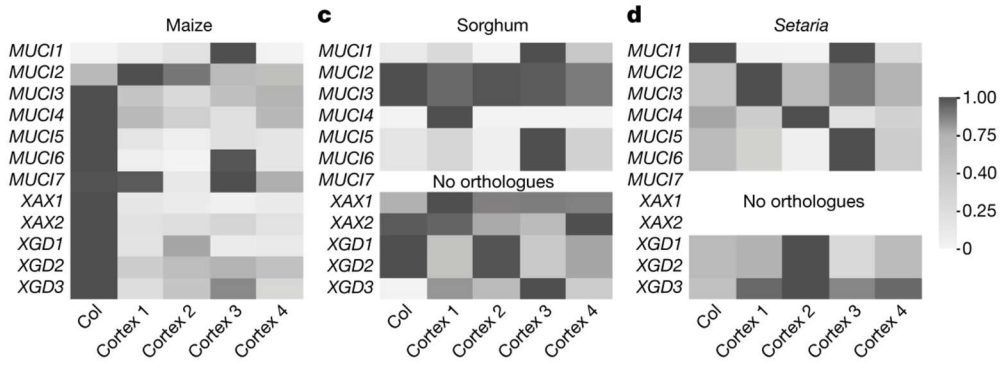


Figure 9. Heat maps showing expression of mucilage genes in maize (b), sorghum (c) and Setaria (d) columella cells and cortex layers.

4.3 Large-scale single-cell profiling of stem cells uncovers redundant regulators of shoot development and yield trait variation

scRNAseq is an emerging technology that enables the capture and profiling of the transcriptional profile of individual cells. By capturing these individual cells, we can tease out information about their development, regulation, and evolution that is obscured in bulk RNA-seq data. Stem cells are arguably the most important cells in a plant, defining the structures that will produce leaves, flowers, and fruits while also defining the overall architecture of the plant. However, despite their importance, shoot stem cells have not been well profiled using scRNA-seq, as a result of 2 main factors. One, shoot stem cells are very delicate, and the harsh chemical removal of the cell wall often damages these cells more than other, more differentiated cell types. Secondly, these cells are very rare, making up only a tiny niche at the tip of the meristem. Because of this, they are exceedingly difficult to capture via scRNA-seq. To mitigate this, our collaborators developed a protocol to efficiently recover stem cells from shoot meristems in both Arabidopsis and maize. In maize, they utilize fine tissue dissection and optimized protoplasting protocols, while in Arabidopsis they utilize over proliferating inflorescence meristems in addition to the optimized protoplasting protocols (Xu et al. 2024). With the ability to capture rare and critical cells in two key model systems, we sought to profile them and compare the stem cell states between these two species.

Although stem cells are highly conserved across plants, making comparisons between maize and Arabidopsis was a much more challenging scenario. Unlike our previous cross-species comparisons, Arabidopsis and maize are separated by nearly 120 million years of evolution, and maize is a monocot, while Arabidopsis is a eudicot. In addition to challenges caused by potential divergence in the transcriptome across this distance, integration is further complicated by a lack of one-to-one gene pairs. Maize and Arabidopsis share only 3653 (as of OrthoDB V11) one-to-one gene pairs, which are the required input for most integration methods. We attempted to integrate the maize and Arabidopsis single cell data with nearly every available integration algorithm, but none were able to achieve a suitable integration (as evaluated by MetaNeighbor)(Crow et al. 2018; Fischer et al. 2021). To address this, we needed to develop a strategy to expand the one-to-one gene space to allow integration. This strategy is addressed in detail in Chapter 2 of this thesis and was published as a separate paper. Using EPIPHITES, we were able to broaden the one-to-one gene space to include 6,075 genes and were able to successfully integrate our datasets (Fig 10). Once we had successfully integrated the data, we evaluated it using MetaNeighbor, finding that meristem and lateral organ clusters had highly conserved transcriptional profiles across the two species. Once more returning to bulk RNA-seq datasets, we looked for which genes expressed in meristem and determinate lateral organ cell types showed high conservation of coexpression, seeking to identify the key conserved modules. Strikingly, we identified a group of conserved meristem markers from the MiniChromosome maintenance family, which are critical in ear development

in maize and function in the embryo and root meristem in Arabidopsis. That these genes showed high conservation despite expression in different tissues suggests the wholesale co-option of the module in one species to perform a conserved role in different tissues.

Here, we again leveraged both novel scRNA-seq data and existing bulk RNA sequencing based coexpression networks to rapidly generate hypotheses and investigate them in the newly generated data. First, we leveraged bulk RNA sequencing networks to make integration between two very distant plant species. Then, using the integrated data, we confirm conservation before turning back to our bulk coexpression networks for hypothesis generation. Identifying potential areas of deep conservation in the bulk network, we investigate these genes in our new dataset, picking out areas of innovation that are unusual in such a deeply conserved cell type.

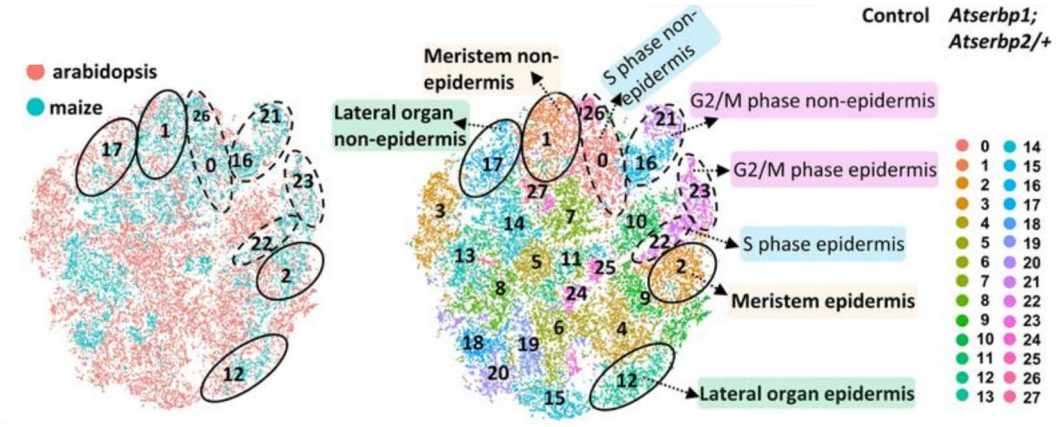


Figure 10. t-Distributed Stochastic Neighbor Embedding (t-SNE) plots shows the integration of arabidopsis and maize shoot tip cells (left panel) resulting in 28 cell clusters (right panel). Dotted ovals indicate dividing cell clusters, solid ovals indicate the conserved meristem and determinate lateral organ clusters

4.4 Convergent evolution of plant prickles by repeated gene co-option over deep time

Similar pressures through natural selection can lead to organisms independently developing similar traits, a process known as convergent evolution. Multiple organisms converge on a similar strategy to address a common pressure. In plants, prickles have evolved multiple times to deter herbivores. However, due to their inconvenience to farmers, prickles are often lost during the domestication of species that have them. This repeated loss enables us to investigate whether the same genetic program is underlying these repeated evolutionary events.

In this study, we worked with our colleagues to better understand if the same genetic modules were involved in the repeated loss of prickles in domesticated solanum (and other) species (Satterlee et al. 2024). To accomplish this, they generated new genomic resources for two Solanum species, *Solanum prinophyllum* and *Solanum cleistogamum*. First, we sought to understand the depth to which the LOG family, the putative gene family responsible for the loss of prickles, was conserved. To do this, we used conservation of coexpression in bulk Arabidopsis and tomato networks to match LOG genes between the species and determine how conserved they were. Initially, it seemed that our results, which showed high conservation of SlyPL (the gene of interest in tomato) to Arabidopsis LOG1, did not match expectations based on the literature, which called SlyPL as a homolog of Arabidopsis LOG8. However, further examination of BLAST results showed that this was likely a miscall, and that the true homolog of SlyPL was

indeed LOG1 (Fig 11). This highlights the power of this framework for capturing conservation and showed that the LOG family genes are highly conserved between Arabidopsis and tomato. Identifying this conservation set us up to compare expression divergence from Arabidopsis to multiple paralogs in the solanum family. We compared LOG1 expression in Arabidopsis, where it is broadly expressed, to the expression of two paralogs (PL and LOG1a) in *Prinophyllum, cleistogamum,* eggplant, and tomato. This comparison showed that PL and LOG1a have both evolved tissue biased expression patterns in the solanum species, with expression being partitioned between the two paralogs (Fig 12). This pattern is common in examples of subfunctionalization, indicating that the gene duplication in Solanum is enabling the specialization of genes in the family.

By leveraging existing high quality coexpression networks from bulk RNA-seq data, we can determine the degree of conservation between gene families even in distant families. When this conservation is high, it indicates that a gene has likely retained a function even across large distances. Here, we use this conservation between two well studied model systems as a jumping off point to investigate conservation in multiple new genetic systems. Additionally, this conservation allowed us to correct a previous missed annotation for homology, and to identify a strong case of gene subfunctionalization.

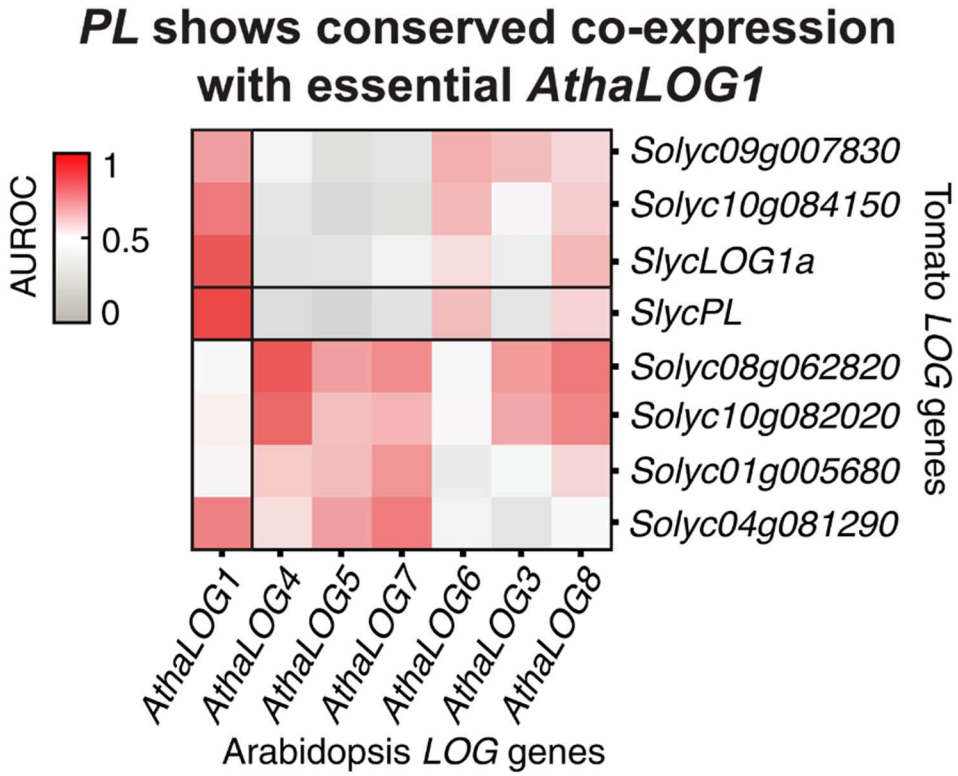


Figure 11. Heatmap depicting the predictability of identifying cross-species coexpressed genes among cross-species pairs of LOG homologs based on their respective coexpression relationships in tomato and arabidopsis. A higher AUROC curve score indicates LOG homologs with increased conservation of their corresponding orthologous coexpressed genes.

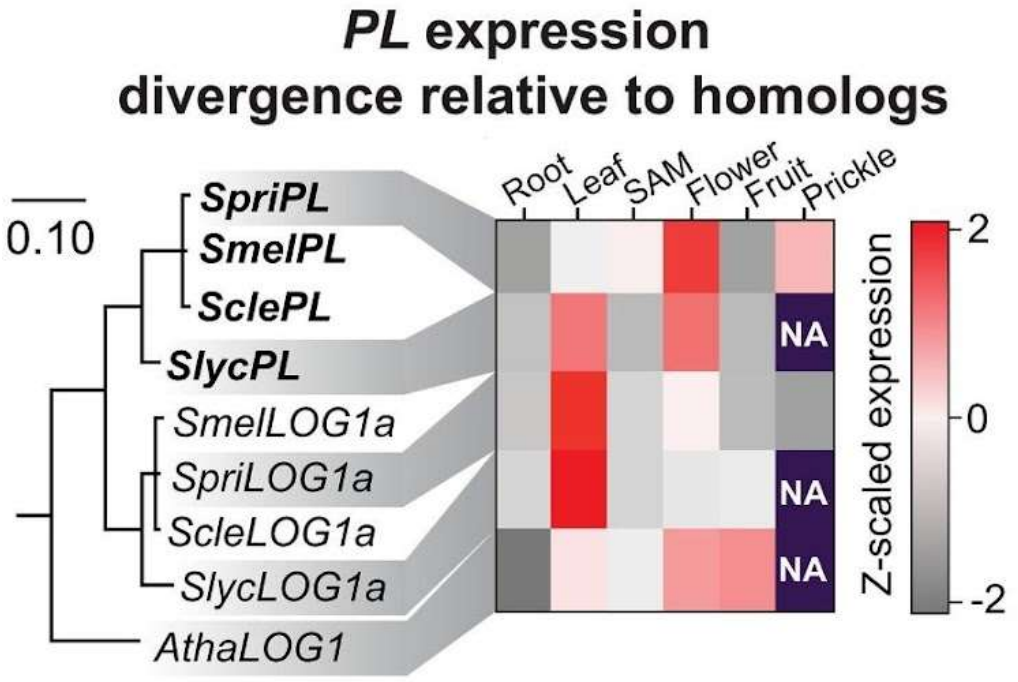


Figure 12. Coding sequence–based maximum likelihood phylogenetic tree of Solanum PL orthologs, their closely related paralog LOG1a, and AthaLOG1 in comparable tissue types. Heatmap shows expression in matched tissues. SAM, shoot apical meristem; NA, not applicable.

Chapter 5

Discussion and Perspectives

This thesis has explored a range of applications of coexpression based analysis to enable and improve integration and analysis of a variety of genomic and transcriptomic data. We lay out two novel frameworks to analyze transcriptomic data, one that improves the integration of scRNA-seq data by expanding the one-to-one gene space, and one that enables us to better capture the underlying coexpression space by merging genes into orthogroups. Additionally, I have highlighted a subset of real world applications of my work, where I used coexpression to better analyze and contextualize novel datasets within the existing world of bulk RNA-seq data. In this chapter, I will explore some of the issues facing the field, limitations of my work, and where I see the field going in the next 5 years. Then, I will summarize our work and its importance to the field more broadly.

5.1 Issues facing the field

5.1.1 Integration vs batch correction

As high throughput sequencing, massively parallel reporter assays, multiplexed CRISPR screens, and other technologies mature, they are increasingly being used to investigate changes across tissues, condition, and even species. These comparisons have been critical for transferring knowledge between species, understanding how genes evolve, and teasing apart the impact of different genetic backgrounds on gene function. However, some high profile initial cross-species comparisons drew surprising conclusions, because of insufficient normalization between samples coming from different species. Unlike most confounds, there is not a clear way to control for species, leading to difficulties with comparison. When comparing samples across species without sufficient normalization, it is easy to conclude that cross-species differences are larger than they truly are, and that information about a gene in one species did not apply to its direct ortholog in another species.

Ultimately, these types of conclusions are usually not borne out, and authors identify that they may have failed to sufficiently batch correct the data. When making cross-species comparisons, it is critical to ensure that results are robust to minor shifts in the integration, and that the integration is sufficient to properly correct for broad, species wide changes in expression while not eliminating important tissue and cell type specific effects that drive phenotypic

differences between species. Unlike over-integration, where a complete lack of structure obscures meaningful differences, comparisons between under-integrated data will identify many spurious differences. These results can give the illusion of success, and identifying under-integration is an important part of any cross-species analysis. Like Goldilocks and the different beds, appropriate integration often involves exploring what the dataset looks like when both over and under integrated, with the goal of finding a level that is just right (or in our case, appropriate and robust).

What makes an integration that is "just right"? While aggressive integration strategies are more likely to erase real biological differences between two samples, insufficient integration or normalization can leave batch effects, impacts from library size, and reference biases in the results. Ideally, integration will retain the inherent structure of the data while removing technical, batch, and non-cell type-specific species effects. Ultimately, there needs to be a careful evaluation of the necessity of integration and a slow and deliberate process to evaluate the data at each step, ensuring changes are robust and sane. First, it is important to determine if a batch correction algorithm is necessary, or if normalization can accomplish the same thing. Additionally, it is important to determine if your question is better served by integration. Pairwise comparisons may benefit less than wider comparisons, and questions that are not tissue specific may be corrected away by integration. Next, it is important to try multiple integration approaches, and to bring multiple approaches through the data analysis pipeline to ensure that the results are robust to integration. As an additional sanity check, you should have

positive and negative "controls" to check for integration failures. Attempting integration of two things that should not integrate, and checking that highly similar things integrate, is important to maintaining data sanity. Finally, it is critical to check the integrity of cell types after integration. Although tempting, relying only on a 2d projection of the data is not sufficient, as methods such as TSNE and UMAP distort the data, often substantially. A common signature of over integrated data is cell types that have all been averaged out to look very similar to each other. Looking at heat maps of expression can identify this and ensure that your data is not over integrated.

5.1.2 Provision of metadata – especially cell types

A key issue limiting the reusability of expensive, meticulously collected scRNA-seq data is the lack of easily available metadata. Currently, a key step in analyzing scRNA-seq data is the manual annotation of cell types based on market sets. This is a fairly labor-intensive process that requires judgement calls and large amounts of subject matter expertise. Unfortunately, when scRNA-seq data is then provided in papers, it is often uploaded to a database such as SRA without metadata labels for cell type available for each unique molecular identifier. This choice dramatically limits the usability of the scRNA-seq data by experts in other species, or by generalists seeking to perform a meta-analysis using the data.

In an ideal world, the upload of read data would be accompanied by a file containing UMI to cell type identifiers. While work is ongoing in the field of

automatic cell type annotation, either based on a foundation model or based on transferring annotation, these models tend to be limited in scope or require data from the same species. For meta-analysis, where there may be only one dataset per species, not providing cell types prevents the reuse of data in a productive way. Beyond just cell type data, several other key pieces of information, such as tissue collected from, plant age, and any stress treatments are key metadata items that may be obscured or lost during the upload process. Improving the provision of metadata would benefit the community and dramatically improve reuse of data that was expensive for the original author, resulting in better science and more citations to those original authors.

5.2 Major limitations and paths forward

5.2.1 Generating sufficient data for bulk RNA networks

EPIPHITES, as well as much of the analysis done on collaborator data, relies on conservation of coexpression, a metric that evaluates how conserved the coexpression partners are of two genes across species. While this metric is very powerful, it does have one major limitation that severely limits its applicability. Calculating coexpression of conservation between two species requires a high quality coexpression network to be available in both species. The construction of these networks is non-trivial, requiring RNA-seq data from dozens of experiments, ideally multiple labs. It is important that these experiments cover as broad a range of tissues and conditions as possible, to accurately capture all coexpression

relationships present. This amount of data is only available for a limited range of species. In plants, we have sufficient data to generate these for only 18 species. When further subset to species that also have high quality orthology information available, we end up limited to only 13 species for which both are available. While these are often the most studied species, plant science is increasingly incorporating new model systems and occurring directly in species of economic impact. Additionally, data becoming publicly available tends to lag its creation by several years, as it is usually only made available when the paper it is generated for is published. Because of these two factors, coexpression networks can end up chasing behind the edge of plant science and may not be available where they are needed most – to better understand new model organisms by transferring knowledge from existing model systems.

5.2.2 Merging large orthology groups washes out signal

While our orthology based approach of merging gene groups to better capture underlying coexpression networks was successful at better identifying deeply conserved families between species, there is one major area where we are aware its performance begins to degrade. In the largest gene families, sub-specialization of genes may be much more advanced, to the point where merging them is no longer an appropriate way to handle the underlying relationship. Instead, these genes may be forming new families of their own, and by merging them

together we may be obscuring these submodules within the family. Unfortunately, it is difficult to distinguish these groups from other large families where most genes are essentially copies of one another, and where merging the genes is both appropriate and improves performance. Because of this, it is important to examine results from large gene families to see if there are submodules. If so, and if a particular family is of interest, it may be possible to subset it into modules based on expression profiles in bulk data prior to merging gene families. These subset gene families are more likely to provide coherent results than averaging several dozen genes.

5.3 Future directions of coexpression analysis

5.3.1 Coexpression networks from single experiments

As demand for analysis of novel model organisms and crop species grows, it will be critical to meet this demand by providing coexpression networks for these species. Fortunately, the recent emergence of scRNA-seq provides a solution to this issue. By treating small groups of ~10 cells as a single sample and averaging all expression within them, we can improve the robustness and minimize dropout/sparsity in single cell data. With enough of these samples, we can build high quality coexpression networks with only a single experiment. These networks are ~80% as good as a gold standard coexpression network built from hundreds of

bulk RNA-seq samples, which is very impressive performance. As scRNA-seq becomes cheaper on a per cell basis, I anticipate that large cell numbers will improve the performance of these networks further. An added benefit of this type of network generation is the ability to generate coexpression networks specific to a condition – for example, an interested scientist could generate a coexpression network of cells under stress to see if there are different coexpression modules that show up. This approach has already been used to elucidate the pathway that produces a paclitaxel precursor, by generating a stressed pacific yew coexpression network from single cell data. This may enable the type of differential network analysis that has not been fruitful with bulk coexpression analysis.

5.3.2 Meta-analysis of single cell experiments

In addition to enabling rapid generation of specific coexpression networks, I believe that the drop in price of generating scRNA-seq data and the proliferation of techniques like pipSeq will facilitate much more meta-analysis of single cell data. Already, a single cell experiment contains so much data that most papers do not fully explore the data. If combined with higher standards for metadata labeling, there will be a tremendous amount of data available to synthesize for analysis. Having abundant labeled single cell data from multiple species will enable extremely broad comparative analysis, tracking the evolution of cell types across species and identifying exactly what modification each species has made vs a likely ancestral state. This degree of labeled data in a wide range of species may also

dramatically improve automated single cell type annotation, improving reproducibility by removing the judgement heavy hand annotation step of pipelines. Finally, more labeled data will enable us to more accurately identify obscured cell types. In many of the scRNA-seq datasets I have analyzed, there have been cryptic variation within canonical cell types such as cortex. I believe that in the next few years, additional scRNA-seq data will allow us to identify the conserved programs underlying this cryptic variation, revealing what these cell types represent.

5.4 Conclusion

My work has focused on comparative analysis of plant species, with a focus on using a coexpression lens to improve cross-species comparisons by analyzing not just genes but also gene modules. It also incorporates emerging technologies, heavily leveraging scRNA-seq to better compare individual cell types between species and understand how gene modules are repurposed between species, resulting in the broad range of phenotypes present in plant species. As technology advances, techniques and computational methods that can best make use of the new data are necessary, and I have attempted to update and build upon existing work to achieve this. As the field marches towards higher depths in single cell and larger panels in spatial data, more data will enable more precise questions. However, it is important to design experiments where these technologies, which

can have substantial drawbacks, are appropriately leveraged. Not all questions are best answered by single cell or spatial technologies, and choosing an appropriate approach is critical to answering interesting questions.

Coexpression networks have provided a key method for answering many different and interesting questions in my work. Conservation of coexpression is uniquely suited to plant species, where tissues can rapidly evolve between species into highly specialized structures with unclear homologies. For example, is the storage tissue of a tuber better compared to storage tissues that are modified stems (bulbs), or to their tissue of origin, roots? Which paralog in the family is the correct homolog to compare a gene to, when both families have multiple genes? With a traditional expression-based comparison, these kinds of questions are critical before you can compare two genes. By instead comparing the conservation of coexpression, we can ask questions about the functional conservation of a gene without having to untie the gordian knot of plant homology. Not only do we not need to identify homologous tissues to interrogate this kind of question, but we also need to sort out the large, complex gene families present in plants. Using conservation of coexpression, we can compare all genes in a family at once, identifying the most similar genes via their expression profile. While this approach is very powerful, it does not tell us where these genes are expressed, what conditions are causing their divergence/conservation, or how their regulation is changing between species. By adding other modalities of data, we can shore up this weakness, providing a very powerful tool for making comparisons between species.

These comparative approaches will be critical as the field of plant science continues to move beyond the initial model systems that many foundational concepts were first understood in and moves towards applications in a broad range of plant species. Here, the massive genetic diversity across plant species is both a blessing and a curse. While this genetic diversity results in the massive range of plant phenotypes harnessed for everything from food, to energy, to construction, this diversity also deeply complicates the transfer of knowledge between plant species. Polyploid genomes, rapid gene duplications, and shifts in both timing and plant architecture complicate cross-species comparisons, making it difficult to stage, align, and compare genes and gene modules. Additionally, molecular techniques often need to be optimized between species, and different preparations make it even more difficult to compare data generated in different species. Despite this, generations of plant scientists have worked to develop deep understandings of the species we rely on, enhancing crop species and selecting for traits that enable us to feed billions of people without needing to increase land use at anywhere near the growth rate of the human population. This work hopes to assist our colleagues in their work, allowing them to more rapidly and easily generate and test hypotheses in their species of interest.

As we generate data in more and more species, we hope that we will begin to uncover the fundamental mechanisms controlling the evolution of plants. We know that repeated cycles of polyploidization and rediploidization have been critical to the radiation and adaptation of plants to different habitats and niches within those habitats. Despite this, we still have a poor understanding of how genes

are repurposed following a whole genome duplication. Are certain genes rapidly predestined for repurposing, while those deemed unnecessary are lost? Or are they retained in a mostly functional state, waiting to step in if the original gene breaks down? Is this process predictable, or does it depend on early, random mutations in the gene regulatory and gene body regions. I hope that this work, as well as other contributions I have made to the field outside the scope of this thesis, will help to answer these questions. By deepening our understanding of the process of evolution, gene sub and neofunctionalization, and species radiation, we can better understand how angiosperms so rapidly dominated our planet. This knowledge is critical not just to sate our curiosity, but to enable us to better control and improve the phenotypes of the many diverse plant species cultivated across the globe.

Bibliography

Aleksander, Suzi A, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, et al. 2023. “The Gene Ontology Knowledgebase in 2023.” *Genetics* 224 (1): iyad031. https://doi.org/10.1093/genetics/iyad031.

Alwine, J. C., D. J. Kemp, and G. R. Stark. 1977. “Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5350–54. https://doi.org/10.1073/pnas.74.12.5350.

Armstrong, Joel, Ian T. Fiddes, Mark Diekhans, and Benedict Paten. 2019. “Whole-Genome Alignment and Comparative Annotation.” *Annual Review of Animal Biosciences* 7 (Volume 7, 2019): 41–64. https://doi.org/10.1146/annurev-animal-020518-115005.

Awika, Joseph M. 2011. “Major Cereal Grains Production and Use around the World.” In *Advances in Cereal Science: Implications to Food Processing and Health Promotion*, 1089:1–13. ACS Symposium Series 1089. American Chemical Society. https://doi.org/10.1021/bk-2011-1089.ch001.

Bainbridge, Matthew N., René L. Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, et al. 2006. "Analysis of the Prostate Cancer Cell Line LNCaP Transcriptome Using a Sequencing-by-Synthesis Approach." *BMC Genomics* 7 (1): 246. https://doi.org/10.1186/1471-2164-7-246.

Ballouz, Sara, Melanie Weber, Paul Pavlidis, and Jesse Gillis. 2017. "EGAD: Ultra-Fast Functional Analysis of Gene Networks." *Bioinformatics* 33 (4): 612–14. https://doi.org/10.1093/bioinformatics/btw695.

Bansal, Payal, Shashi Banga, and S. S. Banga. 2012. "Heterosis as Investigated in Terms of Polyploidy and Genetic Diversity Using Designed Brassica Juncea Amphiploid and Its Progenitor Diploid Species." *PLoS ONE* 7 (2): e29607. https://doi.org/10.1371/journal.pone.0029607.

Baran, Yael, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. 2019. "542524: Analysis of Single-Cell RNA-Seq Data Using K-Nn Graph Partitions." *Genome Biology* 20 (1): 206. https://doi.org/10.1186/s13059-019-1812-2.

Bennett, Richard N., and Roger M. Wallsgrove. 1994. "Secondary Metabolites in Plant Defence Mechanisms." *New Phytologist* 127 (4): 617–33. https://doi.org/10.1111/j.1469-8137.1994.tb02968.x.

Bennetzen, Jeffrey L. 2000. "Comparative Sequence Analysis of Plant Nuclear Genomes: Microcolinearity and Its Many Exceptions." *The Plant Cell* 12 (7): 1021–29. https://doi.org/10.1105/tpc.12.7.1021.

Bergmann, Sven, Jan Ihmels, and Naama Barkai. 2004. "Similarities and Differences in Genome-Wide Expression Data of Six Organisms." *PLoS Biology* 2 (1): e9. https://doi.org/10.1371/journal.pbio.0020009.

Berthelot, Camille, Frédéric Brunet, Domitille Chalopin, Amélie Juanchich, Maria Bernard, Benjamin Noël, Pascal Bento, et al. 2014. "The Rainbow Trout Genome Provides Novel Insights into Evolution after Whole-Genome Duplication in Vertebrates." *Nature Communications* 5 (1): 3657. https://doi.org/10.1038/ncomms4657.

Bevan, Michael W., Richard B. Flavell, and Mary-Dell Chilton. 1983. "A Chimaeric Antibiotic Resistance Gene as a Selectable Marker for Plant Cell Transformation." *Nature* 304 (5922): 184–87. https://doi.org/10.1038/304184a0.

Biase, Fernando H., Xiaoyi Cao, and Sheng Zhong. 2014. "Cell Fate Inclination within 2-Cell and 4-Cell Mouse Embryos Revealed by Single-Cell RNA Sequencing." *Genome Research* 24 (11): 1787–96. https://doi.org/10.1101/gr.177725.114.

Birchler, James A. 2013. "Aneuploidy in Plants and Flies: The Origin of Studies of Genomic Imbalance." *Seminars in Cell & Developmental Biology*, WASP/WAVE proteins: expanding members and functions & The role of ploidy

variation on cellular adaptation, 24 (4): 315–19. https://doi.org/10.1016/j.semcdb.2013.02.004.

Birchler, James A, and Hua Yang. 2022. "The Multiple Fates of Gene Duplications: Deletion, Hypofunctionalization, Subfunctionalization, Neofunctionalization, Dosage Balance Constraints, and Neutral Variation." *The Plant Cell* 34 (7): 2466–74. https://doi.org/10.1093/plcell/koac076.

BLAKESLEE, ALBERT F. 1934. "NEW JIMSON WEEDS FROM OLD CHROMOSOMES." *Journal of Heredity* 25 (3): 81–108. https://doi.org/10.1093/oxfordjournals.jhered.a103898.

Brewer, George J., Charles F. Sing, and E. R. Sears. 1969. "STUDIES OF ISOZYME PATTERNS IN NULLISOMIC-TETRASOMIC COMBINATIONS OF HEXAPLOID WHEAT*." *Proceedings of the National Academy of Sciences of the United States of America* 64 (4): 1224–29.

Calderwood, Alexander, Jo Hepworth, Shannon Woodhouse, Lorelei Bilham, D. Marc Jones, Eleri Tudor, Mubarak Ali, et al. 2021. "Comparative Transcriptomics Reveals Desynchronisation of Gene Expression during the Floral Transition between Arabidopsis and Brassica Rapa Cultivars." *Quantitative Plant Biology* 2 (January):e4. https://doi.org/10.1017/qpb.2021.6.

Chen, Hongyu, Xinxin Yin, Longbiao Guo, Jie Yao, Yiwen Ding, Xiaoxu Xu, Lu Liu, Qian-Hao Zhu, Qinjie Chu, and Longjiang Fan. 2021. "PlantscRNAdb: A

Database for Plant Single-Cell RNA Analysis." *Molecular Plant* 14 (6): 855–57. https://doi.org/10.1016/j.molp.2021.05.002.

Chen, Shuonan, and Jessica C. Mar. 2018. "Evaluating Methods of Inferring Gene Regulatory Networks Highlights Their Lack of Performance for Single Cell Gene Expression Data." *BMC Bioinformatics* 19 (1): 232. https://doi.org/10.1186/s12859-018-2217-z.

Chen, Yanqing, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J. MacNeil, Chunsheng Zhang, et al. 2008. "Variations in DNA Elucidate Molecular Networks That Cause Disease." *Nature* 452 (7186): 429–35. https://doi.org/10.1038/nature06757.

Chen, Zelin, Yoshihiro Omori, Sergey Koren, Takuya Shirokiya, Takuo Kuroda, Atsushi Miyamoto, Hironori Wada, et al. 2019. "De Novo Assembly of the Goldfish (Carassius Auratus) Genome and the Evolution of Genes after Whole-Genome Duplication." *Science Advances* 5 (6): eaav0547. https://doi.org/10.1126/sciadv.aav0547.

Christenhusz, Maarten J. M., and James W. Byng. 2016. "<p>The Number of Known Plants Species in the World and Its Annual Increase</Strong></P>." *Phytotaxa* 261 (3): 201–17. https://doi.org/10.11646/phytotaxa.261.3.1.

Clark, James W., and Philip C.J. Donoghue. 2018. "Whole-Genome Duplication and Plant Macroevolution." *Trends in Plant Science* 23 (10): 933–45. https://doi.org/10.1016/j.tplants.2018.07.006.

Colmer, T. D., and O. Pedersen. 2008. "Oxygen Dynamics in Submerged Rice (Oryza Sativa)." *New Phytologist* 178 (2): 326–34. https://doi.org/10.1111/j.1469-8137.2007.02364.x.

Corneillie, Sander, Nico De Storme, Rebecca Van Acker, Jonatan U. Fangel, Michiel De Bruyne, Riet De Rycke, Danny Geelen, William G. T. Willats, Bartel Vanholme, and Wout Boerjan. 2019. "Polyploidy Affects Plant Growth and Alters Cell Wall Composition1[OPEN]." *Plant Physiology* 179 (1): 74–87. https://doi.org/10.1104/pp.18.00967.

Crepet, William L., and Karl J. Niklas. 2009. "Darwin's Second 'Abominable Mystery': Why Are There so Many Angiosperm Species?" *American Journal of Botany* 96 (1): 366–81. https://doi.org/10.3732/ajb.0800126.

Crow, Megan, Anirban Paul, Sara Ballouz, Z. Josh Huang, and Jesse Gillis. 2018. "Characterizing the Replicability of Cell Types Defined by Single Cell RNA-Sequencing Data Using MetaNeighbor." *Nature Communications* 9 (1): 884. https://doi.org/10.1038/s41467-018-03282-0.

Crow, Megan, Hamsini Suresh, John Lee, and Jesse Gillis. 2022. "Coexpression Reveals Conserved Gene Programs That Co-Vary with Cell Type across

Kingdoms." *Nucleic Acids Research* 50 (8): 4302–14. https://doi.org/10.1093/nar/gkac276.

Darwin, Charles. 1903. *More Letters of Charles Darwin: A Record of His Work in a Series of Hitherto Unpublished Letters*. D. Appleton.

Denyer, Tom, Xiaoli Ma, Simon Klesen, Emanuele Scacchi, Kay Nieselt, and Marja C. P. Timmermans. 2019. "Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Single-Cell RNA Sequencing." *Developmental Cell* 48 (6): 840-852.e5. https://doi.org/10.1016/j.devcel.2019.02.022.

Deynze, Allen Van, Pablo Zamora, Pierre-Marc Delaux, Cristobal Heitmann, Dhileepkumar Jayaraman, Shanmugam Rajasekar, Danielle Graham, et al. 2018. "Nitrogen Fixation in a Landrace of Maize Is Supported by a Mucilage-Associated Diazotrophic Microbiota." *PLOS Biology* 16 (8): e2006352. https://doi.org/10.1371/journal.pbio.2006352.

Dodson, Laura. 2024. "Adoption of Genetically Engineered Crops in the U.S." U.S. Department of Agriculture, Economic Research Service.

Doyle, James A. 1978. "Origin of Angiospers." *Annual Review of Ecology and Systematics* 9:365–92.

Drewes, Gerard, and Tewis Bouwmeester. 2003. "Global Approaches to Protein–Protein Interactions." *Current Opinion in Cell Biology* 15 (2): 199–205. https://doi.org/10.1016/S0955-0674(03)00005-X.

Edger, Patrick P., and J. Chris Pires. 2009. "Gene and Genome Duplications: The Impact of Dosage-Sensitivity on the Fate of Nuclear Genes." *Chromosome Research* 17 (5): 699–717. https://doi.org/10.1007/s10577-009-9055-9.

Edwards, M. D., C. W. Stuber, and J. F. Wendel. 1987. "Molecular-Marker-Facilitated Investigations of Quantitative-Trait Loci in Maize. I. Numbers, Genomic Distribution and Types of Gene Action." *Genetics* 116 (1): 113–25. https://doi.org/10.1093/genetics/116.1.113.

Ehrlich, Paul R. 1983. *The Population Bomb*. Ballantine Books.

Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, and David Botstein. 1998. "Cluster Analysis and Display of Genome-Wide Expression Patterns." *Proceedings of the National Academy of Sciences* 95 (25): 14863–68.

Evenson, R. E., and D. Gollin. 2003. "Assessing the Impact of the Green Revolution, 1960 to 2000." *Science* 300 (5620): 758–62. https://doi.org/10.1126/science.1078710.

Fajkus, Jiří, Eva Sýkorová, and Andrew R. Leitch. 2005. "Telomeres in Evolution and Evolution of Telomeres." *Chromosome Research* 13 (5): 469–79. https://doi.org/10.1007/s10577-005-0997-2.

Fernandez-Cornejo, Jorge, Seth Wechsler, Mike Livingston, and Lorraine Mitchell. 2014. "Genetically Engineered Crops in the United States." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2503388.

Fischer, Stephan, Megan Crow, Benjamin D. Harris, and Jesse Gillis. 2021. "Scaling up Reproducible Research for Single-Cell Transcriptomics Using MetaNeighbor." *Nature Protocols* 16 (8): 4031–67. https://doi.org/10.1038/s41596-021-00575-5.

Flagel, Lex E., and Jonathan F. Wendel. 2009. "Gene Duplication and Evolutionary Novelty in Plants." *New Phytologist* 183 (3): 557–64. https://doi.org/10.1111/j.1469-8137.2009.02923.x.

Folk, Ryan A., Carolina M. Siniscalchi, and Douglas E. Soltis. 2020. "Angiosperms at the Edge: Extremity, Diversity, and Phylogeny." *Plant, Cell & Environment* 43 (12): 2871–93. https://doi.org/10.1111/pce.13887.

Fraley, R T, S G Rogers, R B Horsch, P R Sanders, J S Flick, S P Adams, M L Bittner, et al. 1983. "Expression of Bacterial Genes in Plant Cells." *Proceedings of the National Academy of Sciences of the United States of America* 80 (15): 4803–7.

Galloway, Andrew F., Paul Knox, and Kirsten Krause. 2020. "Sticky Mucilages and Exudates of Plants: Putative Microenvironmental Design Elements with Biotechnological Value." *New Phytologist* 225 (4): 1461–69. https://doi.org/10.1111/nph.16144.

Gharib, Walid H., and Marc Robinson-Rechavi. 2011. "When Orthologs Diverge between Human and Mouse." *Briefings in Bioinformatics* 12 (5): 436–41. https://doi.org/10.1093/bib/bbr031.

Guillotin, Bruno, Ramin Rahni, Michael Passalacqua, Mohammed Ateequr Mohammed, Xiaosa Xu, Sunil Kenchanmane Raju, Carlos Ortiz Ramírez, et al. 2023. "A Pan-Grass Transcriptome Reveals Patterns of Cellular Divergence in Crops." *Nature* 617 (7962): 785–91. https://doi.org/10.1038/s41586-023-06053-0.

Hadebe, S. T., A. T. Modi, and T. Mabhaudhi. 2017. "Drought Tolerance and Water Use of Cereal Crops: A Focus on Sorghum as a Food Security Crop in Sub-Saharan Africa." *Journal of Agronomy and Crop Science* 203 (3): 177–91. https://doi.org/10.1111/jac.12191.

Harris, Benjamin D., Megan Crow, Stephan Fischer, and Jesse Gillis. 2021. "Single-Cell Co-Expression Analysis Reveals That Transcriptional Modules Are Shared across Cell Types in the Brain." *Cell Systems* 12 (7): 748-756.e3. https://doi.org/10.1016/j.cels.2021.04.010.

Hearn, David J., Patrick O'Brien, and Sylvie M. Poulsen. 2018. "Comparative Transcriptomics Reveals Shared Gene Expression Changes during Independent Evolutionary Origins of Stem and Hypocotyl/Root Tubers in Brassica (Brassicaceae)." *PLOS ONE* 13 (6): e0197166. https://doi.org/10.1371/journal.pone.0197166.

Heimberg, Graham, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. 2016. "Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing." *Cell Systems* 2 (4): 239–50. https://doi.org/10.1016/j.cels.2016.04.001.

Heller, Michael J. 2002. "DNA Microarray Technology: Devices, Systems, and Applications." *Annual Review of Biomedical Engineering* 4 (Volume 4, 2002): 129–53. https://doi.org/10.1146/annurev.bioeng.4.020702.153438.

Hendelman, Anat, Sophia Zebell, Daniel Rodriguez-Leal, Noah Dukler, Gina Robitaille, Xuelin Wu, Jamie Kostyun, et al. 2021. "Conserved Pleiotropy of an Ancient Plant Homeobox Gene Uncovered by Cis-Regulatory Dissection." *Cell* 184 (7): 1724-1739.e16. https://doi.org/10.1016/j.cell.2021.02.001.

Herrera-Estrella, Luis, Ann Depicker, Marc Van Montagu, and Jeff Schell. 1983. "Expression of Chimaeric Genes Transferred into Plant Cells Using a Ti-Plasmid-Derived Vector." *Nature* 303 (5914): 209–13. https://doi.org/10.1038/303209a0.

Hickey, Leo J., and James A. Doyle. 1977. "Early Cretaceous Fossil Evidence for Angiosperm Evolution." *The Botanical Review* 43 (1): 3–104. https://doi.org/10.1007/BF02860849.

Hie, Brian, Bryan Bryson, and Bonnie Berger. 2019. "Efficient Integration of Heterogeneous Single-Cell Transcriptomes Using Scanorama." *Nature Biotechnology* 37 (6): 685–91. https://doi.org/10.1038/s41587-019-0113-3.

Hilgenhof, Rebecca, Edeline Gagnon, Sandra Knapp, Xavier Aubriot, Eric J. Tepe, Lynn Bohs, Leandro L. Giacomin, et al. 2023. "Morphological Trait Evolution in Solanum (Solanaceae): Evolutionary Lability of Key Taxonomic Characters." *TAXON* 72 (4): 811–47. https://doi.org/10.1002/tax.12990.

Hoekema, A., P. R. Hirsch, P. J. J. Hooykaas, and R. A. Schilperoort. 1983. "A Binary Plant Vector Strategy Based on Separation of Vir- and T-Region of the Agrobacterium Tumefaciens Ti-Plasmid." *Nature* 303 (5913): 179–80. https://doi.org/10.1038/303179a0.

Holland, Peter W. H. 1999. "Gene Duplication: Past, Present and Future." *Seminars in Cell & Developmental Biology* 10 (5): 541–47. https://doi.org/10.1006/scdb.1999.0335.

Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, et al. 2000. "Functional Discovery via a Compendium of Expression Profiles." *Cell* 102 (1): 109–26. https://doi.org/10.1016/s0092-8674(00)00015-5.

Iancu, Ovidiu D., Sunita Kawane, Daniel Bottomly, Robert Searles, Robert Hitzemann, and Shannon McWeeney. 2012. "Utilizing RNA-Seq Data for de Novo Coexpression Network Inference." *Bioinformatics* 28 (12): 1592–97. https://doi.org/10.1093/bioinformatics/bts245.

Ihmels, Jan, Gilgi Friedlander, Sven Bergmann, Ofer Sarig, Yaniv Ziv, and Naama Barkai. 2002. "Revealing Modular Organization in the Yeast Transcriptional Network." *Nature Genetics* 31 (4): 370–77. https://doi.org/10.1038/ng941.

Isozymes in Plant Genetics and Breeding. 2012. Elsevier.

Jander, Georg, and Carina Barth. 2007. "Tandem Gene Arrays: A Challenge for Functional Genomics." *Trends in Plant Science* 12 (5): 203–10. https://doi.org/10.1016/j.tplants.2007.03.008.

Jean-Baptiste, Ken, José L. McFaline-Figueroa, Cristina M. Alexandre, Michael W. Dorrity, Lauren Saunders, Kerry L. Bubb, Cole Trapnell, Stanley Fields, Christine Queitsch, and Josh T. Cuperus. 2019. "Dynamics of Gene Expression in Single Root Cells of Arabidopsis Thaliana." *The Plant Cell* 31 (5): 993–1011. https://doi.org/10.1105/tpc.18.00785.

Kejnovsky, Eduard, Ilia J. Leitch, and Andrew R. Leitch. 2009. "Contrasting Evolutionary Dynamics between Angiosperm and Mammalian Genomes." *Trends in Ecology & Evolution* 24 (10): 572–82. https://doi.org/10.1016/j.tree.2009.04.010.

Kim, S. K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie, and G. S. Davidson. 2001. "A Gene Expression Map for Caenorhabditis Elegans." *Science (New York, N.Y.)* 293 (5537): 2087–92. https://doi.org/10.1126/science.1061603.

Kozlova, Liudmila V., Alsu R. Nazipova, Oleg V. Gorshkov, Anna A. Petrova, and Tatyana A. Gorshkova. 2020. "Elongating Maize Root: Zone-Specific Combinations of Polysaccharides from Type I and Type II Primary Cell Walls." *Scientific Reports* 10 (1): 10956. https://doi.org/10.1038/s41598-020-67782-0.

Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. "OrthoDB V10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs." *Nucleic Acids Research* 47 (D1): D807–11.

Kumar, Sudhir, Michael Suleski, Jack M Craig, Adrienne E Kasprowicz, Maxwell Sanderford, Michael Li, Glen Stecher, and S Blair Hedges. 2022. "TimeTree 5: An Expanded Resource for Species Divergence Times." *Molecular Biology and Evolution* 39 (8): msac174. https://doi.org/10.1093/molbev/msac174.

Kuznetsov, Dmitry, Fredrik Tegenfeldt, Mosè Manni, Mathieu Seppey, Matthew Berkeley, Evgenia V Kriventseva, and Evgeny M Zdobnov. 2023. "OrthoDB V11: Annotation of Orthologs in the Widest Sampling of Organismal Diversity." *Nucleic Acids Research* 51 (D1): D445–51. https://doi.org/10.1093/nar/gkac998.

Lashkari, Deval A., Joseph L. DeRisi, John H. McCusker, Allen F. Namath, Cristl Gentile, Seung Y. Hwang, Patrick O. Brown, and Ronald W. Davis. 1997. "Yeast Microarrays for Genome Wide Parallel Genetic and Gene Expression Analysis." *Proceedings of the National Academy of Sciences of the United States of America* 94 (24): 13057–62.

Lee, Homin K., Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. 2004. "Coexpression Analysis of Human Genes Across Many Microarray Data Sets." *Genome Research* 14 (6): 1085–94. https://doi.org/10.1101/gr.1910904.

Lee, John, Manthan Shah, Sara Ballouz, Megan Crow, and Jesse Gillis. 2020. "CoCoCoNet: Conserved and Comparative Co-Expression across a Diverse Set of Species." *Nucleic Acids Research* 48 (W1): W566–71. https://doi.org/10.1093/nar/gkaa348.

Li, Xuhui, Xiangbo Zhang, Shuai Gao, Fangqing Cui, Weiwei Chen, Lina Fan, and Yongwen Qi. 2022. "Single-Cell RNA Sequencing Reveals the Landscape of Maize Root Tips and Assists in Identification of Cell Type-Specific Nitrate-Response Genes." *The Crop Journal* 10 (6): 1589–1600. https://doi.org/10.1016/j.cj.2022.02.004.

Lisch, Damon. 2012. "Regulation of Transposable Elements in Maize." *Current Opinion in Plant Biology* 15 (5): 511–16. https://doi.org/10.1016/j.pbi.2012.07.001.

———. 2013. "How Important Are Transposons for Plant Evolution?" *Nature Reviews Genetics* 14 (1): 49–61. https://doi.org/10.1038/nrg3374.

Liu, Qing, Zhe Liang, Dan Feng, Sanjie Jiang, Yifan Wang, Zhuoying Du, Ruoxi Li, et al. 2021. "Transcriptional Landscape of Rice Roots at the Single-Cell Resolution." *Molecular Plant* 14 (3): 384–94. https://doi.org/10.1016/j.molp.2020.12.014.

Lotfollahi, Mohammad, F. Alexander Wolf, and Fabian J. Theis. 2019. "scGen Predicts Single-Cell Perturbation Responses." *Nature Methods* 16 (8): 715–21. https://doi.org/10.1038/s41592-019-0494-8.

Macqueen, Daniel J., and Ian A. Johnston. 2014. "A Well-Constrained Estimate for the Timing of the Salmonid Whole Genome Duplication Reveals Major Decoupling from Species Diversification." *Proceedings of the Royal Society B: Biological Sciences* 281 (1778): 20132881. https://doi.org/10.1098/rspb.2013.2881.

Madlung, A. 2013. "Polyploidy and Its Effect on Evolutionary Success: Old Questions Revisited with New Tools." *Heredity* 110 (2): 99–104. https://doi.org/10.1038/hdy.2012.79.

Malthus, Thomas Robert. 1809. *An Essay on the Principle of Population, as It Affects the Future Improvement of Society*.

Marlétaz, Ferdinand, Panos N. Firbas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry, Christopher D. R. Wyatt, et al. 2018. "Amphioxus Functional Genomics and the Origins of Vertebrate Gene Regulation." *Nature* 564 (7734): 64–70. https://doi.org/10.1038/s41586-018-0734-6.

Messeder, João Vitor S., Tomás A. Carlo, Guojin Zhang, Juan David Tovar, César Arana, Jie Huang, Chien-Hsun Huang, and Hong Ma. 2024. "A Highly Resolved Nuclear Phylogeny Uncovers Strong Phylogenetic Conservatism and Correlated Evolution of Fruit Color and Size in Solanum L." *New Phytologist* 243 (2): 765–80. https://doi.org/10.1111/nph.19849.

Mitreiter, Simon, and Tamara Gigolashvili. 2021. “Regulation of Glucosinolate Biosynthesis.” *Journal of Experimental Botany* 72 (1): 70–91. https://doi.org/10.1093/jxb/eraa479.

Morabito, Samuel, Fairlie Reese, Negin Rahimzadeh, Emily Miyoshi, and Vivek Swarup. 2023. “hdWGCNA Identifies Co-Expression Networks in High-Dimensional Transcriptomics Data.” *Cell Reports Methods* 3 (6): 100498. https://doi.org/10.1016/j.crmeth.2023.100498.

Noort, Vera van, Berend Snel, and Martijn A. Huynen. 2003. “Predicting Gene Function by Conserved Co-Expression.” *Trends in Genetics* 19 (5): 238–42. https://doi.org/10.1016/S0168-9525(03)00056-8.

Oliphant, Adam, Prasad Thenkabail, and Pardhasaradhi Teluguntla. 2022. “Global Food-Security-Support-Analysis Data at 30-m Resolution (GFSAD30) Cropland-Extent Products—Download Analysis.” 2022–1001. *Open-File Report*. U.S. Geological Survey. https://doi.org/10.3133/ofr20221001.

Olsen, Thale Kristin, and Ninib Baryawno. 2018. “Introduction to Single-Cell RNA Sequencing.” *Current Protocols in Molecular Biology* 122 (1): e57. https://doi.org/10.1002/cpmb.57.

Panchy, Nicholas, Melissa Lehti-Shiu, and Shin-Han Shiu. 2016. “Evolution of Gene Duplication in Plants.” *Plant Physiology* 171 (4): 2294–2316. https://doi.org/10.1104/pp.16.00523.

Pickett, F B, and D R Meeks-Wagner. 1995. "Seeing Double: Appreciating Genetic Redundancy." *The Plant Cell* 7 (9): 1347–56. https://doi.org/10.1105/tpc.7.9.1347.

Plant Cell Atlas Consortium, Suryatapa Ghosh Jha, Alexander T Borowsky, Benjamin J Cole, Noah Fahlgren, Andrew Farmer, Shao-shan Carol Huang, et al. 2021. "Vision, Challenges and Opportunities for a Plant Cell Atlas." Edited by Peter Rodgers and Elsa Loissel. *eLife* 10 (September):e66877. https://doi.org/10.7554/eLife.66877.

Ren, Ren, Haifeng Wang, Chunce Guo, Ning Zhang, Liping Zeng, Yamao Chen, Hong Ma, and Ji Qi. 2018. "Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms." *Molecular Plant*, Genome Biology, 11 (3): 414–28. https://doi.org/10.1016/j.molp.2018.01.002.

Rensing, Stefan A., Daniel Lang, Andreas D. Zimmer, Astrid Terry, Asaf Salamov, Harris Shapiro, Tomoaki Nishiyama, et al. 2008. "The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants." *Science* 319 (5859): 64–69. https://doi.org/10.1126/science.1150646.

Ryu, Kook Hui, Ling Huang, Hyun Min Kang, and John Schiefelbein. 2019. "Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells." *Plant Physiology* 179 (4): 1444–56. https://doi.org/10.1104/pp.18.01482.

SATINA, SOPIIIA, A. F. BLAKESLEE, and AMOS G. AVERY. 1937. "BALANCED AND UNBALANCED HAPLOIDS IN DATURA." *Journal of Heredity* 28 (6): 193–202. https://doi.org/10.1093/oxfordjournals.jhered.a104360.

Satterlee, James W., David Alonso, Pietro Gramazio, Katharine M. Jenike, Jia He, Andrea Arrones, Gloria Villanueva, et al. 2024. "Convergent Evolution of Plant Prickles by Repeated Gene Co-Option over Deep Time." *Science* 385 (6708): eado1663. https://doi.org/10.1126/science.ado1663.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science (New York, N.Y.)* 270 (5235): 467–70. https://doi.org/10.1126/science.270.5235.467.

Schuster, Christoph, Alexander Gabel, Hajk-Georg Drost, Ivo Grosse, Ottoline Leyser, and Elliot M. Meyerowitz. 2024. "Rapid Evolution of Gene Expression Patterns in Flowering Plants." bioRxiv. https://doi.org/10.1101/2024.07.08.602577.

Sears, E. R. 1953. "Nullisomic Analysis in Common Wheat." *The American Naturalist* 87 (835): 245–52. https://doi.org/10.1086/281780.

Session, Adam M., Yoshinobu Uno, Taejoon Kwon, Jarrod A. Chapman, Atsushi Toyoda, Shuji Takahashi, Akimasa Fukui, et al. 2016. "Genome Evolution in the Allotetraploid Frog Xenopus Laevis." *Nature* 538 (7625): 336–43. https://doi.org/10.1038/nature19840.

Shulse, Christine N., Benjamin J. Cole, Doina Ciobanu, Junyan Lin, Yuko Yoshinaga, Mona Gouran, Gina M. Turco, et al. 2019. "High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types." *Cell Reports* 27 (7): 2241-2247.e4. https://doi.org/10.1016/j.celrep.2019.04.054.

Song, Yuyao, Yanhui Hu, Julian Dow, Norbert Perrimon, and Irene Papatheodorou. 2024. "ScGOclust: Leveraging Gene Ontology to Compare Cell Types across Distant Species Using scRNA-Seq Data." bioRxiv. https://doi.org/10.1101/2024.01.09.574675.

Sree, K. Sowjanya, Sailendharan Sudakaran, and Klaus-J. Appenroth. 2015. "How Fast Can Angiosperms Grow? Species and Clonal Diversity of Growth Rates in the Genus Wolffia (Lemnaceae)." *Acta Physiologiae Plantarum* 37 (10): 204. https://doi.org/10.1007/s11738-015-1951-3.

Stark, Rory, Marta Grzelak, and James Hadfield. 2019. "RNA Sequencing: The Teenage Years." *Nature Reviews Genetics* 20 (11): 631–56. https://doi.org/10.1038/s41576-019-0150-2.

Stuart, Joshua M., Eran Segal, Daphne Koller, and Stuart K. Kim. 2003. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules." *Science* 302 (5643): 249–55. https://doi.org/10.1126/science.1087447.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and

Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177 (7): 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Su, Andrew I., Michael P. Cooke, Keith A. Ching, Yaron Hakak, John R. Walker, Tim Wiltshire, Anthony P. Orth, et al. 2002. “Large-Scale Analysis of the Human and Mouse Transcriptomes.” *Proceedings of the National Academy of Sciences* 99 (7): 4465–70. https://doi.org/10.1073/pnas.012025199.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. “mRNA-Seq Whole-Transcriptome Analysis of a Single Cell.” *Nature Methods* 6 (5): 377–82. https://doi.org/10.1038/nmeth.1315.

Tarashansky, Alexander J, Jacob M Musser, Margarita Khariton, Pengyang Li, Detlev Arendt, Stephen R Quake, and Bo Wang. 2021. “Mapping Single-Cell Atlases throughout Metazoa Unravels Cell Type Evolution.” Edited by Alex K Shalek and Naama Barkai. *eLife* 10 (May):e66747. https://doi.org/10.7554/eLife.66747.

Teoh, Eng Soon. 2015. “Secondary Metabolites of Plants.” *Medicinal Orchids of Asia*, November, 59–73. https://doi.org/10.1007/978-3-319-24274-3_5.

Walker, M.G. 2001. “Pharmaceutical Target Identification by Gene Expression Analysis.” *Mini Reviews in Medicinal Chemistry* 1 (2): 197–205. https://doi.org/10.2174/1389557013407034.

Wang, Xingang, Lyndsey Aguirre, Daniel Rodríguez-Leal, Anat Hendelman, Matthias Benoit, and Zachary B. Lippman. 2021. "Dissecting Cis-Regulatory Control of Quantitative Trait Variation in a Plant Stem Cell Circuit." *Nature Plants* 7 (4): 419–27.

Wang, Xuran, David Choi, and Kathryn Roeder. 2021. "Constructing Local Cell-Specific Networks from Single-Cell Data." *Proceedings of the National Academy of Sciences* 118 (51): e2113178118. https://doi.org/10.1073/pnas.2113178118.

Waters, David A., Geoffrey E. Burrows, and John D. I. Harper. 2010. "Eucalyptus Regnans (Myrtaceae): A Fire-Sensitive Eucalypt with a Resprouter Epicormic Structure." *American Journal of Botany* 97 (4): 545–56. https://doi.org/10.3732/ajb.0900158.

Wetterstrand, Kris A. 2023. "DNA Sequencing Costs: Data." DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). May 2023. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Witter, Martha S., and Gerald D. Carr. 1988. "ADAPTIVE RADIATION AND GENETIC DIFFERENTIATION IN THE HAWAIIAN SILVERSWORD ALLIANCE (COMPOSITAE: MADIINAE)." *Evolution* 42 (6): 1278–87. https://doi.org/10.1111/j.1558-5646.1988.tb04187.x.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1): 15. https://doi.org/10.1186/s13059-017-1382-0.

Xu, Xiaosa, Michael Passalacqua, Brian Rice, Edgar Demesa-Arevalo, Mikiko Kojima, Yumiko Takebayashi, Benjamin Harris, et al. 2024. "Large-Scale Single-Cell Profiling of Stem Cells Uncovers Redundant Regulators of Shoot Development and Yield Trait Variation." bioRxiv. https://doi.org/10.1101/2024.03.04.583414.

Yang, Ning, Yuebin Wang, Xiangguo Liu, Minliang Jin, Miguel Vallebueno-Estrada, Erin Calfee, Lu Chen, et al. 2023. "Two Teosintes Made Modern Maize." *Science* 382 (6674): eadg8940. https://doi.org/10.1126/science.adg8940.

Zhang, Bin, and Steve Horvath. 2005. "A General Framework for Weighted Gene Co-Expression Network Analysis." *Statistical Applications in Genetics and Molecular Biology* 4 (1). https://doi.org/10.2202/1544-6115.1128.

Zhang, Yang, Daniel W. Ngu, Daniel Carvalho, Zhikai Liang, Yumou Qiu, Rebecca L. Roston, and James C. Schnable. 2017. "Differentially Regulated Orthologs in Sorghum and the Subgenomes of Maize." *The Plant Cell* 29 (8): 1938–51. https://doi.org/10.1105/tpc.17.00354.

Zhao, Tao, and M. Eric Schranz. 2019. "Network-Based Microsynteny Analysis Identifies Major Differences and Genomic Outliers in Mammalian and

Angiosperm Genomes." *Proceedings of the National Academy of Sciences* 116 (6): 2165–74. https://doi.org/10.1073/pnas.1801757116.

Zhuang, Lei, and Haoran Zhang. 2021. "Utilizing Cross-Species Co-Cultures for Discovery of Novel Natural Products." *Current Opinion in Biotechnology*, Chemical Biotechnology ● Pharmaceutical Biotechnology, 69 (June):252–62. https://doi.org/10.1016/j.copbio.2021.01.023.