GENOMICS OF HYBRIDIZATION
Insights into the Ecology and Evolution of Polyploid Plants through Network Analysis

JOSEPH P. GALLAGHER, a CORRINNE E. GROVER, a GUANJING HU and JONATHAN F. WENDEL
Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA

Abstract

Polyploidy is a widespread phenomenon throughout eukaryotes, with important ecological and evolutionary consequences. Although genes operate as components of complex pathways and networks, polyploid changes in genes and gene expression have typically been evaluated as either individual genes or as a part of broad-scale analyses. Network analysis has been fruitful in associating genomic and other ‘omic’-based changes with phenotype for many systems. In polyploid species, network analysis has the potential not only to facilitate a better understanding of the complex ‘omic’ underpinnings of phenotypic and ecological traits common to polyploidy, but also to provide novel insight into the interaction among duplicated genes and genomes. This adds perspective to the global patterns of expression (and other ‘omic’) change that accompany polyploidy and to the patterns of recruitment and/or loss of genes following polyploidization. While network analysis in polyploid species faces challenges common to other analyses of duplicated genomes, present technologies combined with thoughtful experimental design provide a powerful system to explore polyploid evolution. Here, we demonstrate the utility and potential of network analysis to questions pertaining to polyploidy with an example involving evolution of the transgressively superior cotton fibres found in polyploid Gossypium hirsutum. By combining network analysis with prior knowledge, we provide further insights into the role of profilins in fibre domestication and exemplify the potential for network analysis in polyploid species.

Keywords: co-expression, gene expression, gene loss, gene retention, network analysis polyploidy

Received 24 November 2015; revision received 9 March 2016; accepted 22 March 2016

Polyploidy is a widespread phenomenon throughout eukaryotes, with important ecological and evolutionary consequences (Stebbins 1940; Levin 1983; Ramsey & Schemske 1998, 2002; Leitch & Leitch 2008; Van de Peer et al. 2009; Matsushita et al. 2012; Soltis et al. 2014). Although both recent and ancient polyploidy events have been identified in animals and fungi, modern polyploidy is not a particularly active process in these major clades (Muller 1925; Orr 1990; Albertin & Marullo 2012). Conversely, the importance of polyploidy to modern plant species is difficult to overstate. It is now recognized that all flowering plants have experienced multiple rounds of polyploidy at some point in their ancestry (Bowers et al. 2003; Wood et al. 2009; Jiao et al. 2011) and that it remains an active evolutionary and ecological process in many lineages (Ramsey 2011; Soltis & Soltis 2012; Ramsey & Ramsey 2014; McAllister et al. 2015; Wendel 2015).

From a genomic perspective, the myriad consequences of polyploidy for the various ‘omes’ have been broadly evaluated for a number of model angiosperm genera (Wang et al. 2004; Pang et al. 2009; Chelaifa et al. 2010; Bao et al. 2011; Dong & Adams 2011; Ha et al. 2011; Xiong et al. 2011; Chester et al. 2012; Flagel et al. 2012; Koh et al. 2012; Ng et al. 2012; Paterson et al. 2012; Hu et al. 2013, 2015; Page et al. 2016).
2013b; Yoo et al. 2013; Chalhoub et al. 2014; Coate et al. 2014; Guan et al. 2014; International Wheat Genome Sequencing Consortium 2014; Li et al. 2014, 2015; Sehrish et al. 2014; Xu et al. 2014; Zhang et al. 2015b). Among the notable phenomena are intergenomic exchanges between genomes that once were isolated in divergent progenitor diploids (Salmon et al. 2010; Wang & Paterson 2011; Flagel et al. 2012; Guo et al. 2014); biased loss and/or fractionation of duplicated genes (hereafter ‘homoeologs’) (Ozkan et al. 2001; Thomas et al. 2006; Woodhouse et al. 2010; Tang et al. 2012); and a variety of types of changes in gene expression arising with and following polyploidy (Comai et al. 2000; Adams et al. 2003; Wang et al. 2004; Buggs et al. 2009, 2010; Rapp et al. 2009; Chelaifa et al. 2010; Flagel & Wendel 2010; Flagel et al. 2012; Shi et al. 2012; Coate et al. 2014). There also is deep interest in the relationship between these varied phenomena and biological diversification (Ramsey & Schemske 1998; Soltis & Soltis 1999, 2012; Ramsey 2011).

To date, alterations in genes and gene expression arising from polyploidy have been evaluated most commonly on a gene-by-gene basis (e.g. Adams et al. 2003), even when these are aggregated on a genome-wide scale (e.g. Yoo et al. 2013) for purposes of generalization. That is, comparisons have often been made between levels of gene expression between diploids and their polyploid derivatives, for any number of genes within the genomes in question. Because genes do not function in isolation, but as components of complex biological networks, it also is important to study gene expression of networks and pathways, as exemplified for plants in recent studies (Ni et al. 2009; Chang et al. 2010; Bekaert et al. 2011; Coate et al. 2013; Pfeifer et al. 2014). These analyses follow on the pioneering consideration of network responses to polyploidy in yeast (Wagner 2001; Conant & Wolfe 2006; van Hoek & Hogeweg 2009; Qian et al. 2011; Makino & McLyngsith 2012).

Until recently, the analysis of biological networks was unrealistic for most nonmodel organisms; however, technological advances such as RNA-seq and other molecular genetic methodologies, combined with conceptual and computational advances in gene co-expression network reconstruction, have made network analysis an increasingly attractive tool for biological discovery. A gene co-expression network is a map of gene expression correlation among samples; nodes in the network represent genes or transcripts, while connections represent expression relationships (Albert 2005; Langfelder & Horvath 2008). An important advantage of gene co-expression networks is the statistical assignment of genes to clusters, called modules. These co-expression modules become more robust with a greater number and higher diversity of samples across tissues and time points within an organism. With appropriate depth of sampling, comparisons of networks between species, ecotypes or genotypes may reveal changes in modules that explain interesting phenotypic differences. Consequently, we now have the potential to better understand the complex ‘omic’ underpinnings of phenotypic and ecological traits, as exemplified by recent studies of maize domestication (Swanson-Wagner et al. 2012), tomato leaf variability (Ichihashi et al. 2014) and environmental regulation of gene expression in pines (Cañas et al. 2015).

Modern high-throughput techniques also facilitate large-scale identification of other molecular components (e.g. genes, RNAs, proteins, metabolites, epigenetic marks), as well as characterization of their expression patterns and interactions. Consequently, various complex data sets are frequently generated to answer ecological, evolutionary and functional questions. The challenge then becomes how to distill these large multi-dimensional data sets into biologically informative conclusions. Network analysis is designed to capture interactions and dependencies among components, often independent of prior knowledge and the interactions of individual components. For nonmodel species, this is particularly attractive in that it permits genome-scale analysis of ecological or evolutionary traits among conditions, taxa or developmental stages, as well as the identification of gene modules that are likely to be functionally related. Although these correlations are based on the indirect evidence of relationship, strong expression correlation over multiple tissues/stages/etc. is often indicative of functional relationships (e.g. genes activated by promoters with similar regulatory elements). Furthermore, conservation of co-expression patterns among species could indicate the conservation of functional relationships (van Noort et al. 2003).

Analyses of biological networks hold promise for providing insights into the dynamics and resolution of polyploid genomes (Conant & Wolfe 2006; Bekaert et al. 2011; Pfeifer et al. 2014), particularly when network analyses are integrated with other ‘omics’ data sets. Network-based analyses in nonplant models have already demonstrated the utility of gene co-expression networks to reveal functional changes in genes and gene modules (van Noort et al. 2003; Conant & Wolfe 2006). Similar analyses can be used to provide further insight into the age-old question of how two (or more) diverged genomes function in a common nucleus subsequent to merger (at the time of hybridization) and following genome doubling, ultimately enhancing our understanding of polyploid ecology and evolution. We introduce this perspective here, by reviewing the application of network analysis to polyploid research and the layering of additional data types to yield novel insights.
biological insights. We provide a brief empirical example of network analysis from our ongoing work in the cotton genus (*Gossypium*), drawing distinctions between the kinds of insights derived from traditional approaches to gene expression analysis and those derived from network tools.

**Polyploidy and the prospects of gene network analysis**

Network analysis in polyploid species has the capability of providing new perspectives into the interaction among duplicated genes and genomes and the changes that accompany polyploidy. These include patterns of genome-wide gene loss/retention and gene expression changes across the transcriptome. Ultimately, this analysis could disentangle some of the underlying forces that govern polyploid evolution and ecology. Here, we illustrate how network analysis could expand our current knowledge on polyploidy, focusing on gene retention, gene expression alterations and phenotypic changes.

**Gene retention and loss following polyploidy**

One outstanding question concerns the dynamics of duplicate gene loss vs. retention. This question has been considered from several perspectives, including selection related to broad functional categories (De Smet et al. 2013), gene dosage demands (Birchler & Veitia 2012) and preferential retention of homoeologs from the less fractionated of two genomes (Schnable et al. 2011). Each of these possibilities could be enlightened by a deeper consideration of genic interactions in biological pathways and networks, which necessarily entail a number of related functional or mechanistic constraints. Broad characterizations of the types of genes preferentially retained in duplicate, such as transcription and signalling-related genes in *Arabidopsis* (Blanc & Wolfe 2004; Seoighe & Gehring 2004; Maere et al. 2005), structural genes in the Compositae (Barker et al. 2008), or those returned to singleton status, such as photosynthesis-related genes (De Smet et al. 2013), provide evolutionary clues into the dynamics of duplicate gene expression and retention following polyploidy, but of necessity ‘paint with a broad brush’. It seems likely that the compression of complex information into generalized categories conceals interactions among genes that influence retention and loss. We note that some explanations for patterns of duplicate gene loss invoke interaction among genes as important for gene retention or loss. The gene balance hypothesis, for example, posits that proteins that contribute to a multiunit complex are selectively maintained in the correct ratios, as imbalance among components of multiprotein complexes could be deleterious (Papp et al. 2003; Birchler et al. 2005; Birchler & Veitia 2007, 2012; Conant et al. 2014). This concept has been extended to polyploid species, where the entire genic complement has been duplicated; that is, duplicate gene retention may be influenced by stoichiometric or ‘balance’ considerations for the duplicated members of multiprotein complexes. The application of the gene balance hypothesis, however, requires knowledge of interactions among the encoded proteins. For example, Makino & McLysaght (2012) showed that, in human, yeast and *Arabidopsis*, there is an enrichment of genes that have protein interactions with other genes in the same conserved block, as opposed to protein interactions across blocks. Their proximity and interaction suggest that these enriched genes might be coregulated, and are therefore in keeping with the gene balance hypothesis.

An added layer of complexity arises when the duplicated genes are nonidentical, as is the case for many or most genes in allopolyploids, polyploids derived from two divergent genomes. In this case, there may be suboptimal interactions among the protein products from the newly combined genomes, leading to fitness differentials among multi-subunit proteins derived from homo-genomic vs. hetero-genomic protein assembly. In principle, this may result in preferential removal or silencing of specific homoeologs or even a cascade of gene loss sufficient to reduce maladaptive interactions and maintain stoichiometric (gene) balance (Chang et al. 2010; Birchler & Veitia 2012). Insights into these possibilities are best facilitated when genes and their protein products are viewed in the context of their functional connections, or networks. An example of such an approach is provided by Bekaert et al. (2011), who used the *Arabidopsis thaliana* metabolic network, a map of functional interactions, to look at gene retention and loss following ancient polyploidy events; one of their findings was that genes retained in duplicate from the most recent whole-genome duplication were clustered in the network, that is interacting with each other, as would be expected according to the gene balance hypothesis.

**Duplicate gene expression after polyploidy**

One of the important realizations about polyploids is that gene expression is often massively altered relative to diploid progenitors (see Introduction). While gene retention is fundamentally binary (retained or not, the latter including the full spectrum of silencing mutations), gene expression is quantitative. Moreover, the aggregate expression of both homoeologs is of interest, not just the expression of individual members of each duplicate gene pair. In this respect, duplicate gene expression may variably be transgressive relative to the progenitor diploids, additive (or average), or in some sense
mimic only one of the two parents (Yoo et al. 2014). Broad patterns of expression alteration may hint at mechanistic causes of these changes, given predicted functional information of up- or downregulated genes, homoeolog usage, etc.; however, here again information regarding interactions among genes holds promise for providing insight into mechanisms of expression alteration.

Classically, the causes of duplicate gene retention have entailed some form of ‘subfunctionalization’ or ‘neofunctionalization’ (Force et al. 1999; Lynch & Force 2000), although these terms are not mutually exclusive nor wholly sufficient (Conant et al. 2014). Neofunctionalization traditionally invokes a novel and presumably adaptive function for one homoeolog arising post-duplication, while the other homoeolog(s) maintains its ancestral function (Stephens 1951; Ohno 1970), whereas subfunctionalization traditionally occurs as regulatory divergence, such that the ancestral aggregate expression becomes partitioned among homoeologs in the relevant tissues and/or stages (Force et al. 1999; Lynch & Force 2000). In the context of polyploidy, an important early observation was that of Adams et al. (2003), who discovered reciprocal silencing of alternate homoeologs in various tissues from tetraploid cotton, demonstrating that subfunctionalization may be rapidly established following polyploidization. Similar work has since been conducted on a genome-wide scale for a number of species (Duarte et al. 2006; Roulin et al. 2013; Hughes et al. 2014; Renny-Byfield et al. 2014); in some cases, the data are more suggestive of neofunctionalization, although sub- and neofunctionalization may be hard to disentangle in many cases (Conant & Wolfe 2008; Conant et al. 2014).

In the context of the present study the connectivity of individual homoeologs of different gene pairs may be an important consideration, one that only becomes evident using network approaches. For example, Pfeifer et al. (2014) generated a co-expression network for bread wheat, grouping all homoeologs into a single gene node to construct the network, and subsequently evaluating the contribution of each homoeolog to the expression pool for various cell types and developmental stages. Interestingly, they found certain network modules exhibit unbalanced expression bias, that is an overall favouring of expression of homoeologs derived from one parent (Grover et al. 2012), which could be associated with function (based on Gene Ontology categorization) and tissue type. This observation, that functional modules may become biased towards a specific homoeologous genome, has broad implications for understanding the evolution of polyploid species. There may be phenotypic consequences that arise from preferential utilization of functionally related genes derived from only one homoeologous genome (e.g. phenotypic similarity to one parent for a given trait). This preferential utilization or functional differentiation of individual homoeologous networks may be revealed by conducting separate network analyses for each constituent set of homoeologs, rather than on data where gene expression for each duplicate gene pair has been lumped prior to analysis. One can envision that such analyses might reveal several phenomena, including: (i) an overall bias towards one parental genome in a reconstructed gene network; (ii) subfunctionalization of networks, such that the alternate homoeologous networks are used in different tissues/stages, in full or in part; and (iii) network neofunctionalization (all, or some) whereby one set of homoeologs from the same progenitor genome participates in a separate, novel, pathway. Phenomena such as these have been described in Saccharomyces cerevisiae, where co-expression networks appear to have partitioned for ancient sets of paralogs (Conant & Wolfe 2006), and in the neoallopolyploid plant Arabidopsis suecica, where genes with more highly correlated expression were more often derived from the same parent of origin (Chang et al. 2010).

Yet another mechanism that could modulate polyploid gene expression is epigenetic regulation (Madlung & Wendel 2013). Several high-throughput methods (e.g. bisulfite sequencing, ChIP-seq, DNase I hypersensitivity assays) extend the convenience of next-generation sequencing to the epigenetic arena and permit genome-wide surveys of DNA methylation, histone modifications and chromatin state, all of which have the potential to affect gene expression (Deal & Henikoff 2011; Meyer 2011; Furey 2012; Tsompana & Buck 2014; Zhang et al. 2014). Similarly, small RNAs (e.g. siRNAs and miRNAs), which also function in expression regulation, are becoming targets of next-generation sequencing as well (Abrout et al. 2012; Gong et al. 2013; Li et al. 2014). When these features differ between parental genomes, expression of each homoeolog in the polyploid can diverge from that of their parents (Chelaifa et al. 2010; Buggs et al. 2011; Xu et al. 2014). Gene co-expression networks that show altered expression patterns between parental species and polyploids may be caused by these epigenetic effects.

Each of the foregoing aspects of duplicate gene expression feed into the short- and long-term patterns of gene loss and retention following the origin of a polyploid species or clade. A somewhat mysterious observation in this respect is the phenomenon of ‘biased fractionation’ (Woodhouse et al. 2010), whereby a polyploid lineage exhibits biased loss of genes from one progenitor genome as it returns to a diploid state. First demonstrated in Arabidopsis (Thomas et al. 2006), biased fractionation has been observed for a number of other taxa (Schnable et al. 2011; Tang et al. 2012; Renny-
Byfield et al. 2015), suggesting that this may be a common outcome of polyploidy. The causes underlying biased fractionation are not well-understood, but one can imagine how the interplay of phenomena such as gene dosage, maladaptive interactions, biased or novel expression patterns, and others introduced above may affect long-term gene retention bias. One might envision how an enhanced understanding of ‘genes in context’, that is, of regulatory interactions and of biological pathways and networks, may help elucidate the neutral and selective forces that govern loss and retention, and thereby affect phenotypes of relevance to the ecology and evolution of polyploids.

Evolutionary and ecological changes following polyploidy

While the genetic changes listed above are interesting in their own right, the ultimate goal of these studies is to discover the underlying changes that lead to ecologically or evolutionarily important phenotypic properties. A number of traits have been associated with polyploidy in plants, such as enlarged cell size, larger overall organism size and delayed development (i.e. the gigas effect); exploitation of new ecological niches; and physiological or biochemical novelty (Levin 1983; Flgel & Wendel 2009; Ramsey & Ramsey 2014; Soltis et al. 2014). For example, McIntyre (2012) showed that, for a set of Claytonia species of different ploidy levels, polyploids occupied a different niche compared with diploids. In Achillea borealis, Ramsey (2011) performed field transplant experiments showing that hexaploids had a distinct advantage over tetraploids in the xeric dune habitats in which they occur. Coate et al. (2013) showed an increased photoprotective capacity in Glycine dolichocarpa compared with its diploid progenitors, as well as identifying the genetic components that made that increased photoprotection possible. Importantly, these polyploidy-derived traits that allow for changes in habitat are the emergent consequences of the induced changes to different ‘omes’ and how they interact with the environment. These traits are not likely due to changes in the sequence or expression of a single gene or pair of homoeologs, but due to changes in a suite of genes and their interactions. The real goal of applying network analysis is to discover how all of these pieces interact to produce the interesting and unique phenotypes we see in polyploid plants.

Adding power to polyploid networks with prior knowledge and other data types

Polyploidization is accompanied by substantial rewiring of biological networks (De Smet & Van de Peer 2012). While gene co-expression networks can illuminate these changes, the quality of the reconstructed network and the resulting inferences depends both upon the data used to generate the network and the specifics of the biological questions being asked. When considering the underlying transcriptional correlates of a particular phenotypic difference (e.g. fruit size, drought tolerance) between two genotypes, ecotypes or species, it is a common practice to conduct RNA-seq expression profiling experiments, and then subject the transcriptional data to differential expression analysis. Extending these analyses to co-expression networks can allow for further inference (see above, Fig. 1); however, more extensive and thoughtful sampling and the integration of prior knowledge or other data types can increase the inferential power of these networks even further.

The nuances behind network analysis design have been extensively discussed (Albert 2005; Horvath 2009; Krouk et al. 2013; Mitra et al. 2013). In general terms, the power of co-expression network analysis to provide insight depends largely on the question, on the extent of sampling, and on the amount of prior knowledge available (Krouk et al. 2013). A recent example from Helianthus exemplifies what can be learned from appropriately tailoring the experimental data to the biological question. Marchand et al. (2014) utilized gene expression data in H. annuus under nine hormonal treatments from seven time points to build a gene regulatory network (GRN) for drought stress in sunflowers, with a focus on an informed set of candidate genes. From this analysis, they: (i) uncovered hub genes for the drought stress GRN; (ii) discovered a role for nitrate transporters in regulating transpiration; and (iii) connected the abscisic acid-dependent and abscisic acid-independent pathways. Through this targeted approach, Marchand et al. were able to infer causal relationships involved in transcriptional regulation, in addition to co-expression patterns. This perspective sets the stage for further analyses, such as evaluating gene differentiation among Helianthus species and cultivars with various adaptations to drought stress. This approach can be applied to a wide range of experiments utilizing gene expression networks.

Gene co-expression networks are relatively straightforward to generate, and with the addition of prior knowledge and proper sampling, have successfully identified gene-to-gene connections related to phenotype and function. These co-expression relationships, however, reflect the results of a series of direct molecular interactions, for example protein–protein interaction, protein–DNA interaction, membership in metabolic networks and small RNA regulatory interactions. For example, consider a hypothetical protein–DNA interaction that results in joint upregulation of a set of genes. This coregulation may be due to the involvement of these genes in a multimeric complex, and hence direct
coregulation of these components by the protein–DNA interaction complex; alternatively, this interaction may only be made possible when the right multimeric complex is present, which itself promotes co-upregulation of the whole pathway. Supplementing gene co-expression networks with this kind of information leads to more informative network reconstructions better capable of discerning the underlying biological interactions (Li & Jackson 2015; however, see Bloom & Adami 2003 for an example of the caveats of integrating data sets).

Protein–protein interaction networks have been used to consider the consequences of gene and genome duplications, most notably in Arabidopsis thaliana, where the effects of both small-scale and whole-genome duplications were evaluated during network construction (Arabidopsis Interactome Mapping Consortium 2011). The Arabidopsis Interactome Mapping Consortium found that shared protein–protein interactions between paralogs decreased rapidly for duplicates derived from both small-scale and genome-wide duplications, followed by a period of much slower decrease. This trend was mirrored by the divergence in protein sequence, suggesting that, following duplication, the rate of protein evolution and the maintenance of interacting partners are connected. Similarly, specific protein–DNA binding interactions may be assessed (e.g. ChIP-chip or ChIP-seq; Heyndrickx et al. 2014) and layered on gene co-expression networks (Angelini & Costa 2014). Other, less specific assays (e.g. DNase I hypersensitivity assay) can also provide information with respect to the protein-bound regions of the genome (e.g. Zhang et al. 2012), which can subsequently be used to inform more specific protein binding assays (Zhu et al. 2015). Biologically, these protein–DNA interactions may include different forms of transcriptional regulation, such as enrichment or depletion of histones or transcription factor binding sites; in the context of allopolyploidy, differences in protein–DNA interactions between duplicate genomes could contribute to novel phenotypes. Ha et al. (2011), for example, used ChIP-seq to show that homoeologous gene expression patterns were altered via histone modification differences in Arabidopsis. Similar experiments in other polyploid species could help determine the underpinnings of their unique phenotypes or adaptations.

Most sophisticated network analyses have been conducted in model organisms, such as yeast and human, where multiple, layered data types (e.g. RNA-seq plus epigenetic surveys, phenotypes, transcription factor binding sites, etc.) have led to increased understanding of gene regulatory network analyses (Madan Babu & Teichmann 2003; Gao et al. 2004; Zhu et al. 2008; Cookson et al. 2009; Mason et al. 2009; Bocklandt et al. 2011; Langfelder et al. 2012). In yeast, for example, integration of genotypic, expression, protein–protein interaction and transcription factor binding data led to the development of an extremely well-refined and useful gene regulatory network (Zhu et al. 2008). Through integration of well-constructed gene co-expression networks with prior knowledge and these other forms of interaction data, we may best address questions central to understanding the ecological success of polyploid lineages and their specific adaptations.

© 2016 John Wiley & Sons Ltd
Empirical example from *Gossypium*

Network analyses have clear advantages in model systems because of the availability of genomic and other ‘omic’ resources and established experimental methods. In plants, the best developed models for most molecular biological experiments are *Arabidopsis*, rice and several other species with well-developed transformation technologies. These species also have the distinct, bioinformatic advantage of diploidy; as with many other types of analyses, direct application of established protocols and pipelines to polyploid species presents complications arising from redundant genomes, particularly from a bioinformatics perspective. As these challenges become addressed (Mithani et al. 2013; Page et al. 2013a, 2014; Duchemin et al. 2015; Page & Udall 2015), we can now envision experiments that will inform us about network processes in polyploid species.

As an example, we present here an application relevant to the question of the evolution of fibre development and morphology in polyploid *Gossypium hirsutum*, the species that provides most of the world’s cotton crop. Phenotypically, fibres from modern cultivated cotton are vastly improved relative to those of their wild progenitors (Fig. 2), having longer, stronger and finer fibre as a consequence of several millennia of strong directional selection under domestication. Morphologically, ‘fibres’ are single-celled epidermal trichomes arising from the surface of the ovule, already evident on the day the flower opens and thereafter growing rapidly over a period of a couple of months through the stages of primary and secondary wall synthesis, followed by maturation and programmed cell death (Haigler et al. 2012). In our earlier work, we showed that the transcriptome of these cells is highly dynamic and that domestication has led to a massive transcriptomic and proteomic rewiring (Chaudhary et al. 2008; Hovav et al. 2008a,b; Rapp et al. 2009, 2010; Bao et al. 2011; Hu et al. 2013, 2014). One insight emerging from this body of work is that the profilin gene family contains six members (PRF1–PRF6), five of which are upregulated in fibres early in development in modern cotton species relative to its wild progenitor (Bao et al. 2011). Moreover, this same upregulation was shown to have been repeatedly and unknowingly selected under domestication in three independently domesticated cottons. Subsequent differential gene expression analysis of two fibre developmental stages in multiple wild and domesticated accessions suggested that PRF1, PRF3 and PRF4 may be key players in the difference between wild and domestication cotton fibre development (Yoo & Wendel 2014). As profilin plays an important role in actin polymerization, it is reasonable to propose that proteins relevant to cytoskeletal behaviour were targeted by humans during domestication and crop improvement (Bao et al. 2011; Yoo & Wendel 2014). These observations raise many questions; for example, have the profilin genes themselves been the targets of selection, or have their upstream regulators? Are all three important in specifying the domesticated fibre phenotype, or is it one gene that encodes the key player, with others passively co-upregulated? What other genes are up- or downregulated, or coexpressed, during fibre development that may be related to PRF1, PRF3 and PRF4?

Both differential gene expression (DGE) and co-expression network analyses have the potential to increase our understanding of the role of profilin genes in fibre development, although from different angles. To illustrate this, we used RNA-seq data from an ongoing analysis of fibre development for both DGE and network reconstruction (J. Gallagher et al., unpublished). These data consist of four developmental stages (5, 10, 15 and 20 days post-anthesis, or flower opening, hereafter ‘dpa’) from three accessions each of wild and domesticated *G. hirsutum*. Standard DGE analysis (see Methods) of the data revealed 3811 genes upregulated under domestication during fibre development (~domestication + dpa, domesticated−wild, adjusted $P$-value $< 0.05$), including 167 transcription factors and 42 genes with known or suspected involvement in cytoskeleton and cell wall organization (Table 1). The numbers of upregulated and downregulated genes between developmental stages and between wild and domesticated cotton demonstrate several notable results, such as the observation of incredible transcriptomic stasis from 10 to 15 dpa in wild cotton and massive change during the transition to secondary wall synthesis in domesticated *G. hirsutum* between 15 and 20 dpa (Fig. 3; Bao et al. 2011; Rapp et al. 2010; Yoo & Wendel 2014). These results illustrate and confirm the massive scope of transcriptional rewiring that has accompanied the morphological transformations engendered by human selection on this single-celled structure. Of the profilins, we find PRF1 (Gorai.009G028500) is significantly upregulated between domesticated and wild cottons at 10, 15 and 20 dpa (Fig. 4A), in contrast to the earlier quantitative PCR (QPCR) results where PRF1 to PRF5 were all upregulated by domestication (Bao et al. 2011); however, this new result was similar to a previous RNA-seq study where only PRF1 was found to be significantly upregulated by domestication at 10 and 20 dpa (Yoo & Wendel 2014), possibly indicative of the more sensitive nature of QPCR. Previous microarray results also indicate upregulation of a profilin from 7 to 20 dpa (Rapp et al. 2010), which coincides with the upregulation of one of the profilins (Gorai.003G061200) from 15 to 20 dpa in domesticated
cotton (Fig. 4A). The differences in results of these several studies may be due to the greater variance in RNA-seq (vs. QPCR or microarrays), as direct inspection of individual gene expression profiles based on normalized read counts suggests that at least half of the profilins are differentially expressed between wild and domesticated cotton while exhibiting diverse developmental programmes (Fig. 4A). These expression patterns are highly consistent with those of Bao et al. (2011) and further our understanding of variation among profilins, both among gene family members and between wild and domesticated cotton fibres. Without prior interest in the profilin family, however, it is likely the differences in expression of these profilins, and their potential relevance to the cotton fibre developmental programme, would be lost in the morass of differentially expressed gene lists.

The same data were subjected to unsupervised network reconstruction via WGCNA, using default parameters (Langfelder & Horvath 2008). From this analysis, the whole fibre transcriptome was clustered into 48 co-expression modules, with the profilin genes located in five different modules. While similar results could be obtained through more complicated DGE analyses than employed above, network reconstruction (as used here) provides a relatively simple and readily accessible method for distilling complex information into partitioned sets of putatively functionally related genes. Although no expression changes of PRF2 were suggested by DGE results, PRF2 was clustered into the same module (ME) with PRF1 (ME5), which represents genes continuously upregulated during fibre development in domesticated cotton (Fig. 4B). While other modules containing profilin genes (ME1 with PRF5 and Gorai.001G025300, ME2 with PRF4, ME3 with PRF6 and ME12 with PRF3 and Gorai.003G061200; Fig 4B) may indicate coordinated expression changes by domestication limited to a certain developmental stage (such as ME1 and ME2 at 10 dpa), we chose to focus on the module containing PRF1 and PRF2 to exemplify what can be gained through network analysis. The genes comprising this module number 1508, less than half the...
number of upregulated genes in domesticated vs. wild cotton from the DGE gene set. Furthermore, only 1062 are common between the two gene lists, winnowing the DGE candidate gene list by ~2800, but identifying an additional ~500 genes that may be related to PRF1 and its role in fibre development that are overlooked by standard DGE analysis. Included in these ~1500 genes are 71 transcription factors, as well as 16 genes with known or suspected involvement in fibre development (J. Jareczek, unpublished), including 11 that also are present in the DGE gene list. In addition, we extracted putative protein–protein interactions of profilins from a cotton bioinformatics database GraP (Zhang et al. 2015a) and identified a phosphatidylinositol phosphate kinase 10 gene (Gorai.004G153600) that was both bioinformatically predicted to interact with PRF1 and also present in the DGE and co-expression module gene lists. Given that protein kinases are key players in intracellular

### Table 1 Gene families related to cytoskeleton and cell wall organization

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Exemplary references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin-related proteins (ARPs)</td>
<td>Schafer &amp; Schroer (1999)</td>
</tr>
<tr>
<td>Actin capping</td>
<td>Cooper &amp; Schafer (2000) and Hart et al. (2000)</td>
</tr>
<tr>
<td>Callose synthase</td>
<td>Irshad et al. (2008)</td>
</tr>
<tr>
<td>CESA</td>
<td>Arioli et al. (1998) and Betancur et al. (2010)</td>
</tr>
<tr>
<td>CESA-like</td>
<td>Cutler &amp; Somerville (1997)</td>
</tr>
<tr>
<td>COBRA-like</td>
<td>Roudier et al. (2002)</td>
</tr>
<tr>
<td>Coflin/ADF</td>
<td>Carlier et al. (1997)</td>
</tr>
<tr>
<td>eEF1α</td>
<td>Sun et al. (1997)</td>
</tr>
<tr>
<td>EMP24/GP25L</td>
<td>Wang et al. (2008) and Zhou et al. (2010)</td>
</tr>
<tr>
<td>Fimbrin/Plastin</td>
<td>Kovar et al. (2000) and McCurdy &amp; Kim (1998)</td>
</tr>
<tr>
<td>Formin</td>
<td>Deeks et al. (2010)</td>
</tr>
<tr>
<td>Fragmin</td>
<td>Furuhashi &amp; Hatano (1989) and Huang et al. (2011)</td>
</tr>
<tr>
<td>Gelsolin</td>
<td>Huang et al. (2011) and Sun et al. (1999)</td>
</tr>
<tr>
<td>Glycosyl hydrolase</td>
<td>Irshad et al. (2008)</td>
</tr>
<tr>
<td>Kinesin</td>
<td>Oppenheimer et al. (1997)</td>
</tr>
<tr>
<td>Methylesterase inhibitor/pectinesterase</td>
<td>Irshad et al. (2008)</td>
</tr>
<tr>
<td>MS4A/Peroxidase</td>
<td>Irshad et al. (2008) and Shigeto et al. (2013)</td>
</tr>
<tr>
<td>Myosin</td>
<td>Kinkema &amp; Schievelbein (1994), Kinkema et al. (1994) and Yamamoto et al. (1995)</td>
</tr>
<tr>
<td>Profilin</td>
<td>Christensen et al. (1996) and Huang et al. (1996)</td>
</tr>
<tr>
<td>Severin</td>
<td>Brown et al. (1982) and Huang et al. (2011)</td>
</tr>
<tr>
<td>Villin</td>
<td>Khurana et al. (2010) and Tang et al. (2012)</td>
</tr>
<tr>
<td>Xyloglucan transferase</td>
<td>Irshad et al. (2008)</td>
</tr>
</tbody>
</table>
signal transduction, further investigation into this gene
of interest could reveal other functional genes or path-
ways that act in coordination with profilin at important
time points in fibre development.

We note that in the foregoing we considered the
aggregate expression of both duplicated copies (homoe-
ologs) of each profilin gene, in that \textit{G. hirsutum} is an
allopolyploid species. In the context of the present
study, we are especially interested in how network
approaches can illuminate processes in \textit{polyploids}, which
requires analyses wherein each duplicated copy of
each gene is bioinformatically distinguished and
transcriptionally measured. From the standpoint of
understanding polyploidy, we might then ask whether
instances of homoeolog expression bias are randomly
distributed among modules, or whether instead they
are coordinated at this level, which would be suggestive
of selection for network coordination and optimization
(Blanc & Wolfe 2004; Conant & Wolfe 2006; Chang \textit{et al}.
2010). This phenomenon, termed ‘concerted divergence’,
was seen by Blanc & Wolfe (2004) in \textit{Arabidopsis thali-
a} in, where paralogs from an ancient genome duplica-
tion event were found to diverge in gene expression
and form two parallel networks with other paralogs
from the same event. To the extent that concerted divergence of homoeologs exists, it may indicate ecologically or evolutionarily interesting sub- or neofunctionalization of gene regulatory networks following genome merger and doubling. Therefore, evaluating homoeolog expression bias in the context of networks can provide insight into the underlying architecture of relevant phenotypes. For example, by overlaying relative homoeolog expression on our profilin example (Fig. 2), we illustrate patterns of homoeolog expression bias for 5 dpa fibre from wild and domesticated cotton. A first observation is that there exists no detectable overall or module-wide bias in homoeolog usage for either wild or domesticated cotton in 5 dpa fibre. A second observation is that PRF1 homoeolog usage switches from A-homoeolog biased to D-homoeolog biased in wild vs. domesticated cottons, respectively. Third, a similar trend is observed for several genes when looking at the top 10% of nearest neighbours (most closely connected genes) to profilin; that is homoeolog usage bias that was previously near-equal becomes biased towards the D-homoeolog under domestication. This is not true for all nearest neighbours, as one gene coding for a putative LETM1-like protein (Gorai.002G121400) remained slightly A-biased and a second, a putative hydroxyacylglutathione hydrolase gene (Gorai.005G073200), went from near-equal homoeolog usage to being slightly A-biased. At a broader scale, for those genes in the top 25% of nearest neighbours to PRF1, the shift under domestication is more evenly spread between A-bias and D-bias. These observations are notable, as long, spinnable fibres are found only in the A-genome parent of the polyploid species, indicating possible recruitment of the alternative homoeolog for certain pathway segments during the phenotypic transition from the coarse brown fibre of wild cotton to the fine white fibre of domesticated cotton; however, further data are required to evaluate this veracity of this speculation.

The above example is intended to illustrate some possibilities for simple co-expression network analyses in polyploid species, how they can reveal genomic interactions within a polyploid nucleus that would not be evident from more traditional DGE analyses, and in a manner potentially connected to phenotypes of ecological or evolutionary relevance. Additional analytical possibilities will emerge in the future, as the resources and tools are developed. For example, in the above we constructed co-expression networks based on few samples and aggregate expression of homoeologs. As such, we are limited to what essentially are aggregate co-expression relationships of both homoeologs simultaneously, and for only a single cell type (fibres). Extension to multiple tissues may reveal novel network interactions among loci/homoeologs that may be indicative of sub- or neofunctionalization. Furthermore, evidence from other types of analyses (e.g. ChIP-seq, methyl-seq, protein–protein interactions) can be integrated and layered in a network context to reveal multiple subcellular connections that previously were hidden from view, thereby facilitating our understanding of the interactions among genes, their regulation and their evolution in a duplicated context. Such integrative network analysis holds promise for increasing our understanding of the complex foundations of the novel and/or adaptive phenotypes and ecological traits of polyploids (Madan Babu & Teichmann 2003; Gao et al. 2004; Zhu et al. 2008; Cookson et al. 2009; Mason et al. 2009; Bocklandt et al. 2011; Langfelder et al. 2012). The power of network-based analyses will become more fully realized as additional molecular relationships and interdependencies become layered on the expression data, such that a more complete understanding emerges into omics relationships and their connections to ecological and evolutionary traits of interest.

Conclusions

Despite substantial gains in understanding the immediate and long-term consequences of polyploidy, there remain many unanswered questions. Recent research in multiple polyploid systems has described patterns of genomic and transcriptomic change (as well as other measurable omics) on different timescales. These patterns have illuminated many of the phenomena associated with genomic merger (hybridization) and doubling, as well as subsequent cladogenesis and diversification. An exciting prospect is that we are able to reveal the underpinnings of complex phenotypes and ecological traits. As the generation of multiple data types becomes more accessible, we have the opportunity to reveal multiple subcellular connections that previously were hidden from view, thereby facilitating our understanding of how omics changes manifest into evolutionary and ecologically important traits. While massive, integrated data sets (e.g. Zhu et al. 2008; Gerstein et al. 2012) are necessary to truly understand the nuances of interplay among polyploid ‘omes’, the integration of multiple data sets from multiple perspectives will ultimately increase our understanding of the formation of novel and adaptive traits in polyploids, and their evolutionary and ecological significance.

Acknowledgements

The authors thank Josef Jareczek for assistance with identifying cell wall and cytoskeleton genes. The authors are supported by grants from the National Science Foundation and Cotton Incorporated. J. P. Gallagher is supported by National Science Foundation Graduate Research Fellowship DGE1247194.
Methods

RNA was collected from fibres as described previously (Yoo & Wendel 2014). RNA-sequencing was performed on Illumina Hi-Sq2500 at Iowa State University DNA Facility (http://www.dna.iastate.edu/). Reads were mapped to the *Gossypium raimondii* genome (Paterson et al. 2012) using GSNAF with SNP-tolerant mapping (Wu & Nacu 2010; Page & Udall 2015). Differential gene expression analysis was conducted in R software v.3.2.0 (R Core Team 2015) with package DESeq2 (Love et al. 2014). For weighted gene co-expression network analysis, raw read count data were rlog-transformed (regularized logarithm built into DESeq2) and then analysed using the WGCNA package in R with default parameters (Langfelder & Horvath 2008, 2012). The resulting networks were visualized using CYTOSCAPE v3.2.0 (Shannon et al. 2003).

References


Ha M, Ng DWK, Li WH, Chen ZJ (2011) Coordinated histone modifications are associated with gene expression variation within and between species. Genome Research, 21, 590–598.

NETWORK ANALYSIS IN POLYPLOID PLANTS

Kinkema M, Schiefelbein J (1994) A myosin from a higher plant has structural similarities to class V myosins. Journal of Molecular Biology, 239, 591–597.


© 2016 John Wiley & Sons Ltd


Data accessibility

RNA-seq data are archived at Dryad: doi:http://dx.doi.org/10.5061/dryad.256hn and will be made publicly available upon final analysis. Scripts for performing the example analysis are available at https://github.com/Wendellab/MolEcol2016.