Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*

Eli Rodgers-Melnick, Shrinivasrao P. Mane, Palitha Dharmawardhana, et al.

Genome Res. published online October 5, 2011
Access the most recent version at doi:10.1101/gr.125146.111

Contrasting Patterns of Evolution Following Whole Genome versus Tandem Duplication Events in *Populus*

Eli Rodgers-Melnick[1], Shrinivasrao P. Mane[2,7], Palitha Dharmawardhana[3], Gancho T. Slavov[1,4], Oswald R. Crasta[2,5], Steven H. Strauss[3], Amy M. Brunner[6,8], and Stephen P. DiFazio[1,8,9]

[1]Department of Biology, West Virginia University, Morgantown, West Virginia 26506, USA

[2]Virginia Bioinformatics Institute at Virginia Tech, Blacksburg, Virginia 24061, USA

[3]Oregon State University, Department of Forest Ecosystems & Society, Corvallis, OR 97331, USA

[4]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, SY23 3EB, United Kingdom

[5]Chromatin Inc.,3440 S Dearborn St # 280, Chicago, IL 60616-5074, USA

[6]Virginia Tech, Department of Forest Resources & Environmental Conservation, Blacksburg, VA 24061, USA

[7]Dow AgroSciences, 9330 Zionsville Rd., Indianapolis, IN 46268, USA

[8]These authors contributed equally to this work

[9]Author for correspondence:

Stephen P. DiFazio
Department of Biology
West Virginia University
53 Campus Drive
Morgantown, WV 26506-6057
spdifazio@mail.wvu.edu ;
tel (304) 293-5201
fax (304) 293-6363

Abstract

Comparative analysis of multiple angiosperm genomes has implicated gene duplication in the expansion and diversification of many gene families. However, empirical data and theory suggest that whole-genome and small-scale duplication events differ with respect to the types of genes preserved as duplicate pairs. We compared gene duplicates resulting from a recent whole genome duplication to a set of tandemly duplicated genes in the model forest tree *Populus trichocarpa*. We used a combination of microarray expression analyses of a diverse set of tissues and functional annotation to assess factors related to the preservation of duplicate genes of both types. Whole genome duplicates are 700 bp longer and are expressed in 20% more tissues than tandem duplicates. Furthermore, certain functional categories are over-represented in each class of duplicates. In particular, disease resistance genes and receptor-like kinases commonly occur in tandem, but are significantly under-retained following whole genome duplication, while whole genome duplicate pairs are enriched for members of signal transduction cascades and transcription factors. The shape of the distribution of expression divergence for duplicated pairs suggests that nearly half of the whole genome duplicates have diverged in expression by a random degeneration process. The remaining pairs have more conserved gene expression than expected by chance, consistent with a role for selection under the constraints of gene balance. We hypothesize that duplicate gene preservation in *Populus* is driven by a combination of subfunctionalization of duplicate pairs and purifying selection favoring retention of genes encoding proteins with large numbers of interactions.

Introduction

Gene duplication functions as the primary driver of evolutionary novelty within higher eukaryotes (Lynch and Conery 2000; Semon and Wolfe 2007). The recent sequencing of several plant genomes has demonstrated that both whole genome and segmental duplications have played major roles in the expansion of angiosperm gene families (Blanc and Wolfe 2004; Tuskan et al. 2006; Jaillon et al. 2007; Ming et al. 2008; Schnable et al. 2009; Schmutz et al. 2010). Whole genome duplications (WGDs) in particular appear to have occurred recurrently throughout the history of the angiosperm lineage (Blanc and Wolfe 2004; Freeling 2009; Paterson et al. 2010; Jiao et al. 2011). Ancient WGDs in diploid lineages have undergone a fractionation of the polyploid genome, during which chromosomal rearrangements, gene conversions, heightened transposon activity, and epigenetic changes left behind a reduced set of duplicate gene pairs (Gaeta et al. 2006; Chen and Ni 2006; Tate et al. 2009; Wang et al. 2010).

Clusters of duplicated genes have also formed through tandem duplication (TD) processes, which have greatly expanded some gene families, such as the Nucleotide Binding Site - Leucine Rich Repeat (NBS-LRR) subset of plant resistance genes (Meyers et al. 2003; Leister 2004; Kohler et al. 2008). Unequal recombination is thought to be the primary mechanism driving the expansion of these gene clusters (Leister 2004; Babushok et al. 2007; Kane et al. 2010). This occurs when interspersed repetitive elements promote crossing over between nonhomologous segments during meiosis or recombinational repair, resulting in the concomitant introduction of a deletion in one chromosome and an insertion in the other. Tandem duplication can also occur through insertion of retrotransposed genes, although these are thought to insert in a random manner and are often pseudogenized at birth because they lack a promoter and have a processed structure (Zhang et al. 2005; Babushok et al. 2007).

Following duplication, each gene within a paralogous pair may evolve in several ways. For example, it may retain the same set of functions as the ancestral copy (Davis and Petrov 2004), retain only a subset of the original set of functions (subfunctionalization; Force et al. 1999; Lynch and Force 2000), obtain a new function (neofunctionalization), or degrade into a nonfunctional gene (nonfunctionalization; Ohno 1970). Notably, the processes of subfunctionalization and neofunctionalization may not be mutually exclusive. Indeed, the degenerative processes leading to subfunctionalization may act upon silencer elements and thereby promote neofunctionalization (Huminiecki and Wolfe 2004). Seminal theory concerning the fates of duplicate genes predicted that the duplicate copy would be shielded from purifying selection by the ancestral copy, thus promoting pseudogenization in the absence of positive selection for a rare acquired function (Ohno 1970). However, the preservation of large numbers of duplicate genes derived from ancient polyploidy events is difficult to reconcile with a model in which null alleles at duplicate loci are easily fixed by genetic drift (Force et al. 1999). Furthermore, duplicate genes show evidence of purifying selection more consistent with buffering of the ancestral gene function than neofunctionalization (Chapman et al. 2006; Hakes et al. 2007; Warren et al. 2010). Force et al. (1999) reconciled Ohno's original theory with more recent observations by proposing subfunctionalization as a means of preserving duplicate genes in the presence of degenerative mutations targeting both members of a duplicate pair. This hypothesis, known as the duplication-degeneration-complementation (DDC) process, posits that degenerative mutations may knock out independent subfunctions encoded by discrete regulatory elements in duplicate genes, thus requiring preservation of both copies in order to maintain the full complement of ancestral gene functions.

More recent models of duplicate gene evolution suggest that rates of duplicate gene retention vary among protein functional groups. Observations of high retention rates among more connected proteins are consistent with the gene balance hypothesis, which predicts that the fate of duplicate genes largely depends on maintaining a stoichiometric balance among members of macromolecular complexes (Freeling 2006; Birchler and Veitia 2007; Edger and Pires 2009; Birchler and Veitia 2010). This hypothesis also predicts that an increasing number of protein-protein interactions should favor retention of WGD pairs while disfavoring the fixation of TD. Indeed, empirical data in yeast and *Arabidopsis* demonstrate that genes involved in signal transduction and transcription are more likely to be retained following a WGD but less likely to be retained in tandem (Seoighe and Gehring 2004; Davis and Petrov 2005; Maere et al. 2005). Meanwhile, the converse is true for other genes, such as those containing NBS-LRR motifs (Meyers et al. 2003; Leister 2004; Meyers et al. 2005; Zhang et al. 2010).

The availability of whole genome sequence and transcriptome data allows us to test the extent to which natural patterns of retention and divergence conform to the predictions of alternative models of duplicate gene evolution. Under the gene balance hypothesis, we expect that WGD and TD genes should have inverse patterns of retention, with retained WGD genes well conserved and biased toward more central roles in networks (Freeling 2009). Alternatively, a pure subfunctionalization or neofunctionalization process should lead to extensive divergence of expression between duplicates, with retention patterns primarily driven by stochastic processes.

We used the model forest tree *Populus trichocarpa* (Torr. & Gray) to examine the factors involved in the preservation of duplicate gene function and expression. *P. trichocarpa* is an excellent model system for the study of duplicate gene evolution because of the large syntenic

regions conserved from the relatively recent Salicoid WGD that is shared across the Salicaceae, containing nearly 8,000 similarly aged paralogous gene pairs (Sterck et al. 2005; Tuskan et al. 2006; Berlin et al. 2010). Using a combination of coding sequence annotations and microarray expression data, we aimed to accomplish the following: (*i*) identify gene characteristics associated with retention following WGD and TD; (*ii*) delineate the factors associated with diversification of expression patterns for gene pairs resulting from the Salicoid WGD; and (iii) determine the degree to which whole genome patterns of duplicate gene retention and expression conform to the expectations of the DDC hypothesis.

Results

Overview of gene expression and duplications

We studied gene expression across a diverse set of tissues from field-grown *P. trichocarpa* trees, including various vegetative tissues and different stages of reproductive development (Supplemental Table 1). We identified 31,445 genes with significant expression levels in at least one of the 14 tissues analyzed (Supplemental Table 2). The vast majority of genes were expressed in both floral and vegetative tissues. However, 4,306 transcripts were only detected in floral tissues that ranged from early floral development to early and late fruit/seed development stages, and 1,423 transcripts were only detected in vegetative tissues (Supplemental Table 3). Due to source differences between reproductive and vegetative samples, some tissue specificity may be due to sample rather than biological effects. Approximately half of the expressed genes - 15,253 - have paralogs that, based on the fourfold degenerate transversion rate (4DTV) distances and syntenic positions, presumably date to the Salicoid WGD (Tuskan et al. 2006; Tang et al. 2008b). These are hereafter referred to as “retained” Salicoid duplicates. We

also identified 1,196 TD genes with pairwise 4DTV distances comparable to those of the Salicoid duplicates. The sizes of tandem arrays ranged from 2 to 15, with more than half (64%) only containing 2 duplicates.

Factors associated with occurrence of gene duplicates

We used logistic regression to identify significant predictors of occurrence of duplicate genes resulting from WGD and TD. Candidate variables included GO functional categories, breadth of expression, and the genomic length of the gene. Sixteen out of the 18 variables we tested were significant predictors of Salicoid duplicate retention (Nagelkerke pseudo-$r^2 = 0.07$, $P < 2e\text{-}16$), and 12 were significant predictors of TD gene presence (Nagelkerke pseudo-$r^2 = 0.121$, $P < 2e\text{-}16$) (Table 1). Interestingly, 8 of the 11 predictors that were significant in both sets had contrasting effects on the presence of duplicates in either category (Fig. 1).

Gene length was one factor associated with the occurrence of genes in both duplication categories. Gene length was positively associated with the odds of Salicoid retention, while it was negatively associated with the odds of TD occurrence. Furthermore, Salicoid duplicates were significantly longer than all other genes in the genome, while TD genes were significantly shorter than all other genes (Table 2).

Expression breadth (i.e. the fraction of tissues in which significant expression occurs) exhibited a similar pattern to that observed for gene length. Higher expression breadth was associated with greater odds of retention as a Salicoid duplicate pair, and with lower odds of occurrence in a TD (Fig. 1; Tables 1,2). Furthermore, Salicoid duplicates had significantly greater expression breath than all other genes in the genome, while the opposite was true of TD genes (Table 2).

Functional categories of gene duplicates

Genes with transcription factor activity, protein binding activity, kinase activity, phosphatase activity, nucleic acid binding, transporter activity, ligase activity, protease activity, and cation binding activity were associated with significantly higher odds of Salicoid duplicate retention. Conversely, the presence of transmembrane regions, receptor activity, catalytic activity, and stress responsiveness were associated with decreased odds of retention. Remarkably, these same functional categories had the exact opposite effects on odds of occurrence in a TD (Fig. 1).

Three categories – protein binding, transporter activity, and kinase activity – were associated with increased odds of occurrence for both Salicoid and TD genes. However, closer inspection of the composition of these groups revealed substantial differences between tandem and Salicoid duplicates (Supplementary Tables 4-5). Among protein binding genes, TDs were highly enriched for genes with leucine rich repeats (LRR), Ankyrin repeats, Toll/Interleukin-1 Receptor (TIR) domains, and stress responsiveness (Fig. 2A). The LRR and TIR domains are primarily components of plant resistance genes (R-genes) in this data set, which are known to have a tendency to occur in tandem arrays (Meyers et al. 2003; Kohler et al. 2008). Interestingly, proteins containing LRR and TIR domains were under-represented among Salicoid duplicates, while the categories most enriched for Salicoids - RING fingers and DNA-dependent transcriptional regulators – were underrepresented among TDs.

There was also a large discrepancy in the composition of the protein kinase groups between the Salicoid and TD genes (Fig. 2B). Ninety-eight percent of tandemly duplicated protein kinases were annotated for a class of receptor-like kinases (PANTHER acc: PTHR23258), while this

class was underrepresented among Salicoid duplicates. Similarly, 21% of tandem protein kinases, versus 1% of Salicoids, were annotated as S-locus-type glycoproteins, characterized by B-lectin and Pg/Apple/Nematode (PAN) domains. Proteins of this class are primarily known for their roles in self-incompatibility (Bassett et al. 2005; Chen et al. 2006). However, these proteins appear to also have roles in both defense and osmotic stress responses (Bassett et al. 2005; Chen et al. 2006). Salicoid duplicates showed over-retention of a class of protein kinases characterized by the SMART S_TKc annotation (acc. SM00220), which corresponds to the catalytic domain of serine/threonine-specific kinases. Despite its presence within 15% of all protein kinases, no TD genes were annotated for this domain. Closer inspection of this class revealed that most of the proteins were cyclin-dependent or calcium/calmodulin-dependent kinases, suggesting that this annotation primarily identified kinases involved in signal transduction pathways.

Both tandem and Salicoid duplicates were annotated for stress responses. However, genes from the two groups differed with respect to the types of stressors to which they respond (Fig. 2C). Genes involved in defense response and apoptosis were over-represented among TD genes and significantly under-retained among Salicoids. One domain common among plant resistance genes, the TIR domain, was present within 26% of stress related tandem genes but only in 1% of Salicoids. However, genes responding to oxidative stress were over-represented among Salicoid pairs, comprising 27% of all stress-related genes within this group. Also present were proteins with roles in DNA repair (16%), osmotic stress (4%), and heat shock (3%) responses. In contrast, only 8% of tandem stress-related genes were involved in oxidative stress, and 1% had roles in DNA repair. Our list of TD genes did not include any that were annotated for osmotic stress or heat shock responses.

Among TD genes associated with transmembrane regions, there was a significant over-representation of protein kinases, LRR and Ankyrin repeat protein binding domains, and S-locus glycoproteins (Fig. 2D). In turn, Salicoid duplicates showed under-retention of protein kinases, Ankyrin repeat, S-locus glycoproteins, and proteins involved in oxidation-reduction. However, RING finger domains were also significantly over-represented among Salicoid duplicates, concordant with the high retention of this class of proteins among genes annotated for protein binding.

Expression Divergence

There was a complex relationship between expression breadth and the distance between expression patterns for Salicoid duplicates. At low combined breadth (the total fraction of tissues in which at least 1 duplicate has significant expression), the nonparametric correlation behind the conservation calculation inflates the proportion of pairs with conserved expression due to the small number of tissues in which significant expression occurs (Fig. 3). This artifact disappears above a combined breadth of 0.4, after which pairs from the Salicoid duplication show significantly higher conservation than expected by chance. Furthermore, the proportion of Salicoid duplicates showing conserved expression rises substantially beyond a combined breadth of 0.8 (Fig. 3).

We also tested whether the patterns of conservation between gene pairs could be explained by a process of random divergence from an ancestral pattern of expression. Observed expression distances between Salicoid duplicates (mean = 0.674, sd = 0.373) were significantly less than expression distances resulting from a simulation of random divergence processes (mean = 0.972, sd = 0.371). Furthermore, the observed distribution had a significant positive skew (D'Agostino

skewness test: P = 2.074e-14), suggesting that it may be comprised of a mixture of two underlying distributions (Fig. 4). Interestingly, genes annotated for transcription factor and nucleic acid binding activities were significantly overrepresented in the left distribution with more conserved expression patterns, while genes with receptor activity or transmembrane regions were overrepresented in the right distribution with less expression conservation (Fisher's Exact Test, FDR = 0.05; Table 3).

Ratio of nonsynonymous to synonymous substitutions and Expression Imbalance

Salicoid duplicates had a significantly lower ratio of nonsynonymous to synonymous substitutions (d_N/d_S) in coding regions compared to TD genes (Table 2). This was true over a range of d_S values (Supplemental Fig. 1-2). Furthermore, using a forward selection strategy, we found a nonlinear relationship between d_N/d_S and expression distance that depended on the interaction between combined breadth and expression imbalance (r^2=0.188, P < 2.2e-16; Table 4). A nonlinear relationship between d_N/d_S and expression distance was found to be superior to a linear fit using an F-test (F = 7.10, P = 0.007725). The relationship between d_N/d_S and expression distance was most evident at high combined breadth and low expression imbalance (Fig. 5).

Discussion

Our analyses of duplicate pairs resulting from the Salicoid WGD and TD suggest that duplicate gene retention is driven by multiple processes. We have interpreted these results with respect to the DDC model, or subfunctionalization (Force et al. 1999). As expected under a subfunctionalization or neofunctionalization process, we find extensive divergence of expression among retained Salicoid duplicates, which is associated with increased sequence divergence. Furthermore, although our ability to detect subfunctionalization versus neofunctionalization is

limited by the lack of an outgroup, population genetic theory suggests preservation by neofunctionalization should be rare when the effective population size is on the order of that in *Populus* (Lynch et al. 2001). We therefore hypothesize that subfunctionalization is the predominant process for this subset of Salicoid duplicates. However, we also find that many Salicoid duplicates have more conserved expression patterns than expected under random divergence, and that tandemly duplicated genes strongly contrast with genes from the recent whole genome duplication with respect to both structural and expression characteristics. This suggests the degenerative process of subfunctionalization may be counterbalanced by a selective drive to retain highly connected proteins as predicted by the gene balance hypothesis (Birchler et al. 2005; Freeling 2009; Birchler and Veitia 2010).

Salicoid duplicates are longer and more broadly expressed

Salicoid duplicate genes are significantly longer than other genes in the genome, whereas tandem duplicates are significantly shorter. Although increased length may leave these genes more susceptible to loss-of-function mutations, longer genes may also have an enhanced ability to subfunctionalize within the coding region, as the loss of a single exon may not be sufficient to knock out an alternatively spliced gene (Altschmied et al. 2002). The relatively high expression breadth of Salicoid duplicates appears to contradict the expectations of the DDC model, in that each duplicate gene is expected to lose subfunctions following duplication. However, this does not entirely discount the role of degenerative processes because subfunctionalization may not completely partition functions between duplicated genes, and degenerative processes may also lead to neofunctionalization following the loss of silencer elements (Huminiecki and Wolfe 2004). Furthermore, breadth does not account for quantitative subfunctionalization within tissues. Finally, patterns of expression are only a crude approximation of function: duplicate

pairs may have divergent biochemical or physiological roles due to structural differentiation of the proteins even when expressed in the same tissues.

The shorter lengths and lower breadths of TD genes most likely reflect the stochastic processes leading to their birth - – unequal crossing over, transposition, and intrachromosomal recombination - and relatively frequent translocations within the genome (Kong et al. 2007; Freeling 2009; Woodhouse et al. 2010). Each of these may result in the incomplete duplication of genes in the tandem set, and retrotranspositions will entirely remove introns. Similarly, these phenomena will also often fail to copy the complete set of ancestral regulatory regions, thereby producing subfunctionalized genes at birth. Therefore, shorter genes that function in a tissue-specific manner may be more likely to produce a viable copy following the degenerative process of tandem duplication. These genes would also be more likely to pseudogenize following any single duplication event due to their limited number of *cis* regulatory elements. However, the gene balance hypothesis also predicts that TDs of longer genes with more sites of protein-protein interaction should be quickly eliminated due to the deleterious effects of dosage imbalance (Birchler and Veitia 2010). We contend that the higher d_N/d_S observed between TD genes is consistent with weaker purifying selection for the retention of ancestral gene functions. Furthermore, the proportion of TD genes for which there was evidence positive selection (i.e., $d_N/d_S > 1$) was an order of magnitude higher than for Salicoid duplicates (P = 7.75e-10). This would suggest that for TDs, maintenance of the full complement of ancestral gene functions may not be the primary factor favoring duplicate retention, as expected under a DDC process (Force et al. 1999).

Salicoid and tandem genes are enriched for different functional categories

Under the expectations of the DDC model, we did not expect major differences in the functional composition of Salicoid and tandem genes. However, even GO functional categories that were predictors of retention in both duplication classes differed strongly in their specific composition .The low retention of RING fingers among TD genes and their high retention among the Salicoid duplicates may fit the gene balance hypothesis, if these domains primarily mediate interactions within protein complexes. However, LRRs mediate a wide-range of protein-protein interactions (Kobe and Kajava 2001), so their low retention following the Salicoid duplication provides an exception to the predictions of the gene balance hypothesis. The majority of both whole genome and tandemly duplicated LRR genes are annotated for serine/threonine protein kinase activity and primarily belong to the receptor-like kinase (RLK)/Pelle family, which has involvement in both defensive and developmental processes (Shiu et al. 2004; Afzal et al. 2008). While the vast majority of these proteins remain functionally uncharacterized in *Populus trichocarpa*, experimental evidence from rice and *Arabidopsis* suggests that lineage-specific expansions of RLKs in these species were primarily driven by the duplication of defense-related genes (Shiu et al. 2004). The role of RLKs in defense is further supported by data for *Oryza*, *Glycine*, and *Gossypium* that demonstrate a positive correlation between the sizes of the NBS and RLK gene families, both within and among species, with little variation due to ancient polyploidization events (Zhang et al. 2010). Together, this evidence suggests that most LRR proteins evolve rapidly in response to local biotic threats (Mondragon-Palomino 2002; Lehti-Shiu et al. 2009).

Among Salicoid duplicate genes bearing LRRs, we observe an especially strong under-retention of defense-related genes, which contrasts sharply with the pattern of retention among TD genes. The high propensity of R-genes to occur in tandem clusters is well documented in

317 plants, including *Populus trichocarpa* (Leister 2004; Meyers et al. 2005; Tuskan et al. 2006;
318 Kohler et al. 2008), and the discrepancy may be explained by the limited domains of R-gene
319 expression. The tandem stress-related genes in our study have a significantly lower breadth than
320 both the stress-related Salicoid group and the tandem gene average (data not shown). This
321 implies that the loss of subfunctions through a DDC process following the Salicoid duplication
322 would be more likely to completely eliminate expression of these stress-related TD genes,
323 assuming their reduced breadth corresponds to a reduced number of regulatory elements. The
324 gene balance hypothesis also predicts that R-genes should be freed from purifying selection for
325 retention of both duplicates, assuming they are not interaction network hubs. One intriguing
326 possibility is that R-genes may encounter negative selection following WGD because of the
327 trade-off between the increased fitness conferred by resistance when the pathogen is present and
328 decreased fitness in its absence (Tian et al. 2003; Meyers et al. 2005). Thus, retention of the
329 entire complement of functional R-genes during the diploidization process may have had fitness
330 costs in the absence of an onslaught of new pathogens, leading to selection against redundant
331 resistance loci.

332 Lastly, we observed that TD genes tend to display much more extreme patterns of
333 functional overrepresentation than Salicoids. This illustrates an important distinction between the
334 modes of duplication. While the set of Salicoid duplicates reflects the pattern of retention
335 following a single duplication event, the tandemly duplicated genes reflect both the propensity of
336 certain genes to duplicate in tandem and the rate of elimination by selection or drift. In the case
337 of R-genes, stress is known to increase rates of somatic recombination, which can facilitate the
338 birth of new genes through unequal crossing over if it occurs within reproductive cell lineages
339 (Mcdowell and Simon 2006). Moreover, the presence of repeat elements, including the repeated

structures of genes within tandem arrays, can also increase the rate of tandem duplication by promoting unequal crossing over (Jelesko et al. 1999). Indeed, an increased rate of birth may explain how gene families can expand through duplication in the absence of purifying selection due to gene balance constraints or the capacity to subfunctionalize with a limited repertoire of ancestral subfunctions. Similarly, the properties of tandem arrays may provide a mechanism for their rapid deletion following WGD, as these regions are known to have high rates of intrachromosomal recombination (Woodhouse et al. 2010).

Expression divergence consistent with two different patterns

Although the expression patterns of Salicoid duplicates were generally more conserved than expected by chance, we interpreted the skewness of the distribution of expression distances for pairs of Salicoid duplicates as an indication that this distribution might be appropriately modeled as a mixture of two distributions. One interpretation based on the estimated parameters of our mixture model is that approximately half (45%) of the duplicate gene pairs diverge in expression following a pattern consistent with a DDC process, wherein regulatory elements randomly degenerate, eventually leading to complete subfunctionalization (Force et al. 1999). The remaining 55% of duplicate gene pairs appear constrained to maintain some redundancy in their expression patterns. Such a constraint is consistent with the expectations of the gene balance hypothesis, in that the drive to retain post-duplication stoichiometric balance should lead to more conserved patterns of expression (Aury et al. 2006; Birchler and Veitia 2010). This is supported by the observation that the GO category for nucleic acid binding proteins – including highly connected transcription factors and ribosomal proteins – was significantly overrepresented in the left distribution. Moreover, the proportion of well conserved Salicoid duplicates is

positively associated with the combined breadth of the genes, suggesting that more ubiquitous expression of the putative ancestral gene leads to greater conservation of its descendants.

Expression and sequence divergence correlated for broadly expressed genes

Under the expectations of the DDC model, we would predict that the degradation of regulatory elements would occur independently of coding sequence mutations, as the duplicates are assumed to be initially redundant and unconstrained by requirements to maintain the same set of interactions (MacCarthy and Bergman 2007). In contrast, the gene balance hypothesis predicts that the expression patterns of duplicates will be constrained primarily by their protein-protein interactions (Birchler and Veitia 2010). This implies that changes in expression may accompany nonsynonymous mutations, thereby maintaining interactions within an evolving network. Our results suggest that both processes may be affecting Salicoid duplicates.

We found that there was no discernible relationship between expression distance and d_N/d_S for gene pairs with low combined expression breadth and high expression imbalance. However, broadly expressed (i.e., combined breadth above 80%) gene pairs did have a significant nonlinear relationship between expression distance and d_N/d_S. Previous studies in *Populus* revealed that more broadly expressed genes tend to be predominantly *cis* regulated, while *trans* regulation drives more tissue-specific expression (Quesada et al. 2008; Drost et al. 2010), a finding that is broadly consistent with our results. Our data further suggest that expression divergence in broadly-expressed genes may be reflected in coding sequence polymorphisms as well as variation in *cis* regulatory domains in noncoding regions. Results from previous studies indicated that the most broadly expressed genes tend to have the slowest evolutionary rates (Pal et al. 2001; Ettwiller and Veitia 2007). This is consistent with the negative correlation we observed

between combined breadth and d_N/d_S. Moreover, under the constraints of the gene balance hypothesis, we would expect the most highly connected genes to be subject to the greatest selective pressure for maintenance of balanced expression between interacting subunits. Indeed, interacting subunits of macromolecular complexes tend to have positively correlated patterns of expression and evolutionary rates (Ettwiller and Veitia 2007). Therefore, we predict that the strongest relationship between expression distance and d_N/d_S occurs for the most highly connected genes, which are selected for *cis* regulatory elements and protein motifs that allow them to maintain balance within the paleopolyploid protein interaction network. This will be the subject of further investigations in our lab.

Conclusion

We find the pattern of duplicate gene retention following the Salicoid WGD in *Populus* broadly consistent with the predictions of the gene balance hypothesis. Future investigations should use a network-based approach to directly gauge whether the most connected genes are most highly retained following WGD. Approximately half of the Salicoid duplicate gene pairs showed patterns of divergence that suggest many whole genome duplicates are not subject to constraints for maintenance of redundancy. A future analysis of nonconserved noncoding regions would enable us to more accurately determine the role of degenerative processes in the observed expression patterns of duplicate genes.

We believe our findings are relevant to the evolution of duplicate genes across a wide range of higher eukaryotes. Taken together, our findings imply the patterns of retention and functional conservation observed following duplication events are contingent upon both random and selective forces, although one or the other tends to predominate depending upon the type of

duplication and the biochemical function of the gene. This study therefore serves as the groundwork for more detailed studies of the relative roles of neutral processes and natural selection in shaping the functional landscape of the paleopolyploid genome.

Methods

Populus Whole Genome Microarray Experiments

The *Populus* whole genome microarray was constructed by Roche NimbleGen (http://www.nimblegen.com/) to target 55,794 nuclear, 69 chloroplast, and 58 mitochondrial gene models predicted from version 1.1 of the *Populus trichocarpa* genome (Tuskan et al. 2006). The array included three 60mer oligonucleotide probes per gene target, which were evaluated for identity to other non-target gene models using WU-BLASTN as an index of potential cross-hybridization and then further refined based on Nimblegen's design guidelines. Prior to the final design, all non-unique probes were removed along with probes for targets with high identity to transposable elements.

Tissues for microarray hybridizations were obtained from field-grown *Populus trichocarpa* trees near Corvallis, OR, USA except that one root sample was collected from in vitro-grown plants and seeds were germinated in vitro. All tissue was obtained from clone Nisqually-1, except for floral tissues and seeds/seedlings, which were collected from wild *P. trichocarpa* trees. Tissues and abbreviations are described in Supplementary Table 1. RNA isolation, labeling and hybridization were conducted as described in Dharmawardhana et al (Dharmawardhana et al. 2010). Two biological replicates were used for each tissue sample, and three biological replicates were used for the xylem sample.

Collection and normalization of microarray data

The NimbleGen microarray data processing pipeline (NMPP) was used to normalize the data using a two-step normalization procedure (Li et al. 2006; Wang et al. 2006). In the first step, quantile normalization was performed among replicates within each tissue, followed by global normalization to adjust all tissues to a similar baseline. ANOVA was used to identify significant differentially expressed genes between tissues. For each gene, the significance of the differences in the mean of the $\log_2$ of intensities between any two tissues was calculated using t-tests. False discovery rate (FDR) was calculated to correct for multiple testing problem (Benjamini and Hochberg 1995). A gene expression difference with estimated positive (up) or negative (down) fold change at alpha=0.05 and FDR (q)=0.05 was considered significant.

Thresholds for significant expression within the microarray were set using probes for 3,149 transposable elements as negative controls, using the 95th percentile as a cutoff. Tissues were tested for correspondence among replicates (Supplemental Figs. 3-4) and individual replicates were clustered based on Euclidean distances (Supplemental Fig. 5). Tissue samples that showed strong correspondence between replicates and which clustered together were used for subsequent analyses.

Although the microarray was based on gene models from *Populus* version 1.1, the results were ported to *Populus* version 2.0 using an approach based on synteny and reciprocal best hits following BLASTP of the models against one another. Prior to the analysis, tissue replicates were averaged together, and values were averaged over version 1 gene models if ambiguity existed in the version 1 to version 2 mapping. Average expression values were set to zero if they did not exceed the threshold set by the negative controls. Otherwise, they were set to the observed average value minus the tissue threshold. Only gene models with significant expression in at least one tissue were used in subsequent analyses.

Identification of duplicate pairs

Potential duplicate pairs were identified by using BLASTP to compare all version 2 gene models against one another. Models with at least 50% identity over at least half the length of the larger gene were considered potential duplicates. The fourfold degenerate transversion rate (4DTV) was calculated after Smith-Waterman alignment of potential duplicates using an affine gap penalty (gap open = -10, gap extension = -1) and the BLOSUM 60 matrix for scoring. The corresponding CDS sequences were then superimposed onto the alignment, and 4DTV was calculated as the number of transversions within 4-fold degenerate sites divided by the total number of 4-fold degenerate sites (Hellsten et al. 2007).

MCScan was then used to discover intragenomic syntenic blocks (Tang et al. 2008a). Average 4DTV was calculated for syntenic segments by taking the mean across gene pairs found by MCScan in each segment. Potential duplicates on syntenic segments with average 4DTV between 0.08 and 0.15 were considered to be Salicoid duplicates, while gene pairs in these regions with 4DTV greater than 0.2 were filtered out due to concerns that they may have arisen from a more ancient duplication event. Potential duplicates that occurred within 100 kb of each other were defined as tandemly duplicated. In order to limit the set of tandemly duplicated genes to pairs with similar divergence times to the Salicoids, we only selected TDs with 4DTV between 0.06 and 0.16, a range that included most of the Salicoid duplicates but did not include the most recent TDs. Tandemly duplicated genes were permitted to have multiple Salicoid duplicates on a syntenic segment if they conformed to the previously stated criteria in order to avoid biases against tandem duplicates in the Salicoid WGD set.

Functional Annotation

Each version 2 peptide sequence was annotated using 14 applications (BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, patternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther, and Gene3D) in conjunction with InterProScan (Quevillon et al. 2005). The resulting InterPro annotations were then cross-referenced to GO terms using the Gene Ontology SQL database dbxref table (http://www.geneontology.org/GO.database.shtml). Because many of the resulting GO categories were very specific, we identified broader categories by identifying nodes that were descendents of the following GO categories within the Gene Ontology hierarchy: protein binding (GO:0005515), transcription factor activity (GO:0003700), catalytic activity (GO:0003824), receptor activity (GO:0004872), ion channel activity (GO:0005216), stress response (GO:0006950), protein kinase activity (GO:0016301), phosphatase activity (GO:0016791), nucleic acid binding (GO:0003676), transporter activity (GO:0005215), ligase activity (GO:0016874), protease activity (GO:0008233), and cation binding (GO:0043169).

Measures of expression divergence between duplicate pairs

The expression distance for each duplicate pair was defined as $1 - S_{xy}$, where S_{xy} denotes the Spearman correlation coefficient between the tissue expression values of each gene following the normalization procedures described above. Combined breadth refers to the number of tissues in which either member of a duplicate pair has significant expression divided by the total number of tissues, and expression imbalance denotes the maximum breadth of the two duplicate genes divided by the minimum breadth. Additionally, d_N/d_S, the ratio of the nonynonymous substitution rate to the synonymous substitution rate, was calculated for all duplicate pairs using the maximum likelihood method of Yoder and Yang implemented in the codeml module of PAML (Yang 2007).

Statistical Analyses

Logistic regression was used to determine significant predictors of duplicate gene retention. Each gene model with evidence of significant expression was given an indicator variable of 0 or 1 for the presence of a Salicoid duplicate and/or having at least one tandem duplicate, and all tandemly duplicated genes were considered as distinct entities. GO functional categories were given an indicator variable of 0 or 1 depending on absence or presence, respectively. Breadth was defined as the fraction of tissues in which a gene showed significant expression (Huminiecki and Wolfe 2004). Gene length was measured in kilobases between the start and stop codon, including all intron sequence. Logistic regression was then carried out using the generalized linear model, binomial family, logit link within the R programming environment. Initially, all variables were added to the model as main effects, and backward selection was used to choose the best set of predictor variables. The significance of each independent variable was then assessed by resampling 1,000 times with replacement. The amount of variance explained by each model was then quantified using the Nagelkerke pseudo-r^2 measure implemented in the lrm function of the Design package in the R programming language.

The statistical significance of differences among duplicate types for gene length and breadth was also tested using an approximative normal quantile (van der Waerden) test for independence, as implemented in the COIN package for the R programming language (Hothorn et al. 2008).

Prior to the subsequent analyses of individual duplicate pairs, we permitted each gene to be present once if, for example, a given gene had multiple tandem duplicates on the paralogous segment from the Salicoid duplication. In such cases, a single duplicate pair was chosen randomly.

Because the maintenance of networks involves the conservation of both coding sequence and regulatory elements, the extent to which these were associated within Salicoid duplicate pairs was also investigated. Using the expression distance between Salicoid duplicates as the dependent variable, the following model was fit with forward selection, using least-squares regression as implemented in the *lm* function of the R programming language:

$$y = \left(\frac{d_N}{d_S}\right)^2 + \frac{d_N}{d_S} + Combined\ Breadth + Exp.Imbalance + \frac{d_N}{d_S} \times Exp.Imbalance + \frac{d_N}{d_S} \times Combined\ Breadth + \frac{d_N}{d_S} \times Combined\ Breadth \times Exp.Imbalance$$

Combined breadth was defined as the proportion of tissues in which either of the duplicate genes had significant expression, while expression imbalance was defined as the breadth of the more broadly expressed gene divided by the breadth of the more narrowly expressed gene.

Distribution of expression distances under random divergence

An empirical distribution of expression distances was constructed for Salicoid duplicates under a model of divergence consistent with a DDC process. This random model assumed the existence of an ancestral gene for each duplicate pair with expression in each tissue corresponding to the maximum of the descendant's expression levels as well as random divergence from this ancestral expression pattern through purely degenerative processes. For each pair of Salicoid duplicates, a putative ancestral gene expression profile was constructed wherein each tissue was assumed to express the maximum of expression values of the two duplicate genes in that tissue. In order to replicate the observed differences in expression during the simulation, a distribution of expression differences was constructed. The percent difference between the two expression values for each tissue was added to a vector, which was

subsequently divided into 50 bins of equal width, including 100% divergence (loss of expression for one of the duplicates in that tissue). Simulated duplicate genes were then made for each putative ancestral gene, wherein the putative ancestral expression level was assigned to each duplicated gene. These duplicates then underwent simulated divergence, during which one gene would be randomly selected in each tissue to have its expression reduced by a quantity sampled from the distribution of percent differences in the observed data. The expression distance was then calculated for each pair of genes as indicated above.

Modeling of the observed expression distances

The distribution of Spearman correlation coefficients is known to be related to Student's t distribution (Press et al. 1992), which converges to the normal distribution with large N. Because the expression distance measure is has a range between 0 and 2, the observed distribution of expression distances was modeled as a mixture of two truncated normal distributions (Johnson et al. 1994). The Nelder-Mead simplex algorithm provided by the *optim* function in the R programming language was used to maximize the following log-likelihood function for values of the observed spearman expression distances between Salicoid duplicates:

$$\ln\left(L(\pi_1, \mu_1, \sigma_1, \pi_2, \mu_2, \sigma_2 | D)\right) = \sum_{i=1}^{n} \ln(\pi_1 f(x_i, \mu_1, \sigma_1) + \pi_2 f(x_i, \mu_2, \sigma_2))$$

where D denotes the data from the observed distribution of Spearman expression distances, π values denote the mixture proportions of each distribution, and μ and σ denote the means and standard deviations of truncated normal distributions between 0 and 2. Initial parameter values were drawn from uniform distributions, wherein means (μ) were permitted to range between the 5^{th} and 95^{th} percentiles of the observed distribution, standard deviations (σ) were permitted to

range between 1/10 and the entire standard deviation of the observed distribution, and mixture proportions (π) were permitted to range between 0.1 and 0.9 with a sum-to-one constraint. Fitting of the mixture model was performed 100 times until the convergence at a relative tolerance of 1E-10. Final parameters were estimated by using a weighted average where the final likelihoods for each iteration served as weights.

Salicoid gene pairs were assigned to the mixture distributions using likelihood ratios. Gene pairs with a likelihood ratio for a distribution greater than or equal to 3 were assigned to the corresponding distribution. Otherwise, they were not assigned to any distribution. The overrepresentation of functional categories within distributions was then assessed using Fisher's exact test, with multiple testing error controlled using the false discovery rate, as implemented in the q-value package for the R programming language.

Data Access

Microarray data used in this work have been submitted to GEO under accession numbers GSE21481 and GSE21485.

Acknowledgements

We thank Jennifer Hawkins and Phil Turk for helpful comments on the manuscript. This work was supported by the Office of Science (BER), U.S. Department of Energy (DOE), Grants No. DE-FG02-06ER64185 and DE-FG02-07ER64449 and by funding from the BioEnergy Science Center, a U.S. DOE Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

Figure Legends

Figure 1. Log_2 of the exponentiated logistic regression coefficients for categories significant for both retention of Salicoid duplicates and the presence of TDs. The log_2 scales the odds ratios for each category such that those above and below 1 have comparable effects.

Figure 2. Relative frequencies of specific annotations within the broad functional categories of protein binding (A), stress response (B), protein kinase (C), and transmembrane regions (D). Annotations were generated by InterproScan, and those shown were among the most common within each broader functional category. Relative protein frequency refers to the fraction of proteins within the broad category that contain the specific annotation. Error bars indicate 95% confidence intervals generated by 1,000 bootstrap replicates.

Figure 3. The proportion of Salicoid duplicate genes with conserved expression plotted against the combined breadth of the two duplicate genes for the observed and random distributions. Conserved expression was defined as a Spearman expression distance less than 0.3, which corresponded to the lower 5% cutoff of expression distances for permuted genes. Error bars denote the standard error obtained by 1,000 bootstrap replicates.

Figure 4. Line plots of the histograms for the observed distribution of Spearman expression distances (red) and the simulated distribution under the assumption of random divergence (green). The probability density functions for each of the 2 truncated normal mixture model distributions are also shown at their mixture proportions (blue and purple). Parameters for the left mixture distribution were mean (μ_1) = 0.3907, standard deviation (σ_1) = 0.2470, mixture proportion (π_1) = 0.5294. Parameters for the right distribution were μ_2 = 0.9630, σ_2 = 0.2944, and π_2 = 0.4706.

Figure 5. Spearman expression distance plotted against d_N/d_S and conditioned on the combined breadth and the expression imbalance between the two genes within each duplicate pair. The predicted expression distance based on fitting to $d_N/d_S + (d_N/d_S)^2$ is shown in red for each subplot.

604 **Table 1.** Significant predictors for retention of genes in Salicoid and tandem duplicate pairs

	Salicoid Duplicates					TDs				
	β	exp(β)	SE	z	*P*(>|z|)	β	exp(β)	SE	z	*P*(>|z|)
(Intercept)	-0.57	0.57	0.03	-20.71	0.00	-2.30	0.10	0.07	-34.558	0.00
Tandem	-1.35	0.26	0.08	-17.26	0.00	N/A	N/A	N/A	N/A	N/A
Salicoid	N/A	N/A	N/A	N/A	N/A	-1.35	0.26	0.08	-17.22	0.00
Breadth	0.57	1.77	0.04	15.90	0.00	-1.22	0.30	0.10	-12.11	0.00
Gene Length	0.10	1.11	0.01	16.90	0.00	-0.13	0.88	0.02	-6.88	0.00
Protein Binding	0.29	1.34	0.04	6.58	0.00	0.81	2.24	0.10	8.69	0.00
Transmembrane	-0.10	0.90	0.03	-3.83	0.00	0.36	1.43	0.07	5.38	0.00
Transcription Factor	0.66	1.93	0.09	7.18	0.00	-0.81	0.44	0.36	-2.25	0.02
Catalytic	-0.19	0.83	0.03	-6.10	0.00	0.30	1.35	0.07	4.00	0.00
Receptor	-1.49	0.23	0.28	-5.31	0.00	0.58	1.78	0.23	2.46	0.01
Stress	-0.74	0.48	0.09	-8.13	0.00	1.29	3.63	0.12	10.40	0.00
Kinase	0.25	1.28	0.06	4.35	0.00	0.31	1.36	0.13	2.50	0.01
Phosphatase	0.70	2.01	0.15	4.61	0.00	N/A	N/A	N/A	N/A	N/A
Nucleic Acid Binding	0.15	1.16	0.04	3.34	0.00	N/A	N/A	N/A	N/A	N/A
Transporter	0.17	1.19	0.06	2.70	0.00	0.48	1.62	0.14	3.40	0.00
Ligase	0.61	1.84	0.11	5.36	0.00	N/A	N/A	N/A	N/A	N/A
Protease	0.24	1.27	0.09	2.65	0.01	N/A	N/A	N/A	N/A	N/A
Cation Binding	0.13	1.14	0.04	3.33	0.00	-0.23	0.79	0.11	-2.11	0.03

605

606

607 **Table 2.** Mean, median, and standard error for the gene length, expression breadth, and d_N/d_S (of
608 pairs) within the whole genome, among Salicoid duplicates, and among TDs. *P*-values for gene
609 length and breadth are based on van der Waerden tests for independence of gene characteristic
610 and duplicate type.

	Group	Mean	Median	SE	P-value
Gene Length	Whole Genome	2.587 kb	2.007 kb	0.0126	
	Salicoid Duplicates	2.854 kb	2.304 kb	0.0183	0.0000
	TDs	2.126 kb	1.767 kb	0.0471	0.0000
Breadth	Whole Genome	0.5396	0.5714	0.00185	
	Salicoid Duplicates	0.5774	0.6428	0.00264	0.0000
	TDs	0.3868	0.2857	0.00875	0.0000
d_N/d_S	Salicoid Duplicates	0.2728	0.2428	2.29e-5	
	TDs	0.4478	0.40295	6.57e-4	

611

612

Table 3. Proportion of genes annotated with a given functional category in each of the two distributions of the mixture model. *P*-values were obtained using Fisher's Exact test, while the q-value is the false discovery rate (FDR) analogue.

	Mix. Dist. 1	Mix. Dist. 2	p-value	q-value
Protein Binding	0.11	0.11	0.92	0.51
Transmembrane	0.28	0.31	0.01	0.03
Transcription factor	0.04	0.02	0.00	0.00
Catalytic	0.29	0.32	0.07	0.11
Receptor	0.00	0.00	0.02	0.04
Ion	0.00	0.00	0.53	0.38
Stress	0.01	0.01	0.11	0.14
Kinase	0.05	0.05	0.22	0.21
Phosphatase	0.01	0.01	0.55	0.38
Nucleic Acid Binding	0.14	0.10	0.00	0.00
Transporter	0.04	0.04	0.71	0.44
Ligase	0.02	0.02	0.75	0.44
Protease	0.03	0.02	0.21	0.21
Cation binding	0.12	0.13	0.47	0.38

Table 4. Regression table for the relationship between expression distance and d_N/d_S, combined breadth, and expression imbalance. The estimate refers to the linear regression coefficient for each predictor variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5331	0.0481	11.09	0.0000
d_N/d_S	0.2607	0.1829	1.43	0.1540
$(d_N/d_S)^2$	-0.1915	0.0719	-2.66	0.0077
Combined Breadth	-0.2004	0.0511	-3.92	0.0001
Exp. Imbalance	0.0719	0.0032	22.40	0.0000
d_N/d_S:Combined Breadth	0.5982	0.1907	3.14	0.0017
d_N/d_S:Exp. Imbalance	-0.1903	0.0281	-6.78	0.0000
d_N/d_S:Combined Breadth:Exp. Imbalance	0.1324	0.0313	4.23	0.0000

References

Afzal AJ, Wood AJ, and Lightfoot DA. 2008. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol. Plant Microbe In* **21**:507-517

Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, Volff JN, and Schartl M. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics* **161**:259-267

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**:171-178

Babushok DV, Ostertag EM, and Kazazian HH. 2007. Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cell. Mol. Life Sci.* **64**:542-554

Bassett CL, Nickerson ML, Farrell Jr RE, Artlip TS, El Ghaouth A, Wilson CL, and Wisniewski ME. 2005. Characterization of an S-locus receptor protein kinase-like gene from peach. *Tree Phys.* **25**:403-411

Benjamini Y and Hochberg Y. 1995. Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B Met.* **57**:289-300

Berlin S, Lagercrantz U, von Arnold S, Ost T, and Ronnberg-Wastljung AC. 2010. High-density linkage mapping and evolution of paralogs and orthologs in *Salix* and *Populus*. *BMC Genomics* **11**

Birchler JA and Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* **19**:395

Birchler JA and Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* **186**:54-62

Birchler JA, Riddle NC, Auger DL, and Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends Genet.* **21**:219-226

Blanc G and Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**:1667-1678

Chapman BA, Bowers JE, Feltus FA, and Paterson AH. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *P. Natl. Acad. Sci. USA* **103**:2730-2735

Chen X, Shang J, Chen D, Lei C, Zou Y, Zhai W, Liu G, Xu J, Ling Z, Cao G, et al. 2006. A B-lectin receptor kinase gene conferring rice blast resistance. *Plant J.* **46**:794-804

Chen ZJ and Ni Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* **28**:240-252

Davis JC and Petrov DA. 2005. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.* **21**:548

Davis JC and Petrov DA. 2004. Preferential Duplication of Conserved Proteins in Eukaryotic Genomes. *PLOS Biol.* **2**:e55

Dharmawardhana P, Brunner AM, and Strauss SH. 2010. Genome-wide transcriptome analysis of the transition from primary to secondary stem development in Populus trichocarpa. *BMC Genomics* **11**

Drost DR, Benedict CI, Berg A, Novaes E, Novaes CRDB, Yu QB, Dervinis C, Maia JM, Yap J, Miles B, et al. 2010. Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. *P. Natl. Acad. Sci. USA* **107**:8492-8497

Edger PP and Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17**:699-717

Ettwiller L and Veitia RA. 2007. Protein coevolution and isoexpression in yeast macromolecular complexes. *Comp. Funct. Genom.* **DOI: 10.1155/2007/58721**

Force A, Lynch M, Pickett B, Amores A, Yan Y, and Postlthwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531-1545

Freeling M. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* **16**:805-814

Freeling M. 2009. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annu. Rev. Plant Biol.* **60**:433-453

Gaeta R, Pires J, Iniguez-Luy F, Leon E, and Osborn T. 2006. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell* **28**:240-252

Hakes L, Pinney JW, Lovell SC, Oliver SG, and Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* **8**

Hellsten U, Khokha MK, Grammer TC, Harland RM, Richardson P, and Rokhsar DS. 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biol.* **5**

Hothorn T, Hornik K, van de Wiel MAV, and Zeileis A. 2008. Implementing a Class of Permutation Tests: The coin Package. *J. Stat. Softw.* **28**:1-23

690 Huminiecki L and Wolfe KH. 2004. Divergence of spatial gene expression profiles following
691 species-specific gene duplications in human and mouse. *Genome Res.* **14**:1870

692 Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg SÃ,
693 Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral
694 hexaploidization in major angiosperm phyla. *Nature* **449**:463-467

695 Jelesko JG, Harper R, Furuya M, and Gruissem W. 1999. Rare germinal unequal crossing-over
696 leading to recombinant gene formation and gene duplication in Arabidopsis thaliana. *P.*
697 *Natl. Acad. Sci. USA* **96**:10302-10307

698 Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu
699 Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms.
700 *Nature* **473**:97-U113

701 Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions*. Wiley
702 Interscience, New York.

703 Kane J, Freeling M, and Lyons E. 2010. The Evolution of a High Copy Gene Array in
704 Arabidopsis. *J. Mol. Evol.* **70**:531-544

705 Kobe B and Kajava AV. 2001. The leucine-rich repeat as a protein recognition motif. *Curr.*
706 *Opin. Struc. Biol.* **11**:725-732

707 Kohler A, Rinaldi CÃ, Duplessis SÃ, Baucher M, Geelen D, Duchaussoy Fdr, Meyers BC,
708 Boerjan W, and Martin F. 2008. Genome-wide identification of NBS resistance genes in
709 Populus trichocarpa. *Plant Mol. Biol* **66**:619-636

710 Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, and others. 2007. Patterns of gene
711 duplication in the plant SKP1 gene family in angiosperms: evidence for multiple
712 mechanisms of rapid gene birth. *Plant J.* **50**:873-885

713 Lehti-Shiu MD, Zou C, Hanada K, and Shiu SH. 2009. Evolutionary History and Stress
714 Regulation of Plant Receptor-Like Kinase/Pelle Genes. *Plant Phys.* **150**:12-26

715 Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of
716 plant disease resistance genes. *Trends Genet.* **20**:116-122

717 Li L, Wang XF, Stolc V, Li XY, Zhang DF, Su N, Tongprasit W, Li SG, Cheng ZK, Wang J, et
718 al. 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nature*
719 *Genetics* **38**:124-129

720 Lynch M and Conery JS. 2000. The evolutionary fate and consequences of duplicate genes.
721 *Science* **290**:1151

722 Lynch M and Force A. 2000. The probability of duplicate gene preservation by
723 subfunctionalization. *Genetics* **154**:459

Lynch M, O'Hely M, Walsh B, and Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**:1789-1804

MacCarthy T and Bergman A. 2007. The limits of subfunctionalization. *BMC Evol. Biol.* **7**:213

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, and Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *P. Natl. Acad. Sci. USA* **102**:5454-5459

Mcdowell JM and Simon SA. 2006. Recent insights into R gene evolution. *Mol. Plant. Pathol.* **7**:437-448

Meyers B, Kaushik S, and Nandety R. 2005. Evolving disease resistance genes. *Curr. Opin. Plant Biol.* **8**:129-134

Meyers BC, Kozik A, Griego A, Kuang H, and Michelmore RW. 2003. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell* **15**:809

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**:991-9U7

Mondragon-Palomino M. 2002. Patterns of Positive Selection in the Complete NBS-LRR Gene Family of Arabidopsis thaliana. *Genome Res.* **12**:1305-1315

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.

Pal C, Papp B, and Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158**:927-931

Paterson A, Michael F, Tang H, and Xiyin W. 2010. Insights from the Comparison of Plant Genome Sequences. *Annu. Rev. Plant Biol.* **61**:349-372

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK.

Quesada T, Li Z, Dervinis C, Li Y, Bocock PN, Tuskan GA, Casella G, Davis JM, and Kirst M. 2008. Comparative analysis of the transcriptomes of Populus trichocarpa and Arabidopsis thaliana suggests extensive evolution of gene expression regulation in angiosperms. *New Phytol.* **180**:408-420

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, and Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**:W116-W120

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**:178-183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**:1112-1115

Semon M and Wolfe K. 2007. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* **17**:505-512

Seoighe C and Gehring C. 2004. Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. *Trends Genet.* **20**:461-464

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, and Li WH. 2004. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* **16**:1220-1234

Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, and Van de Peer Y. 2005. EST data suggest that poplar is an ancient polyploid. *New Phytol.* **167**:165-170

Tang H, Wang X, Bowers JE, Ming R, Alam M, and Paterson AH. 2008a. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**:1944-1954

Tang HB, Bowers JE, Wang XY, Ming R, Alam M, and Paterson AH. 2008b. Perspective - Synteny and collinearity in plant genomes. *Science* **320**:486-488

Tate JA, Joshi P, Soltis KA, Soltis PS, and Soltis DE. 2009. On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biol.* **9**:80

Tian D, Traw MB, Chen JQ, Kreitman M, and Bergelson J. 2003. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. *Nature* **423**:74-77

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**:1596-1604

Wang XF, He H, Li L, Chen RS, Deng XW, and Li SG. 2006. NMPP: a user-customized NimbleGen microarray data processing pipeline. *Bioinformatics* **22**:2955-2957

Wang Y, Jha AK, Chen R, Doonan JH, and Yang M. 2010. Polyploidy-associated genomic instability in *Arabidopsis thaliana*. *Genesis* **48**:254-263

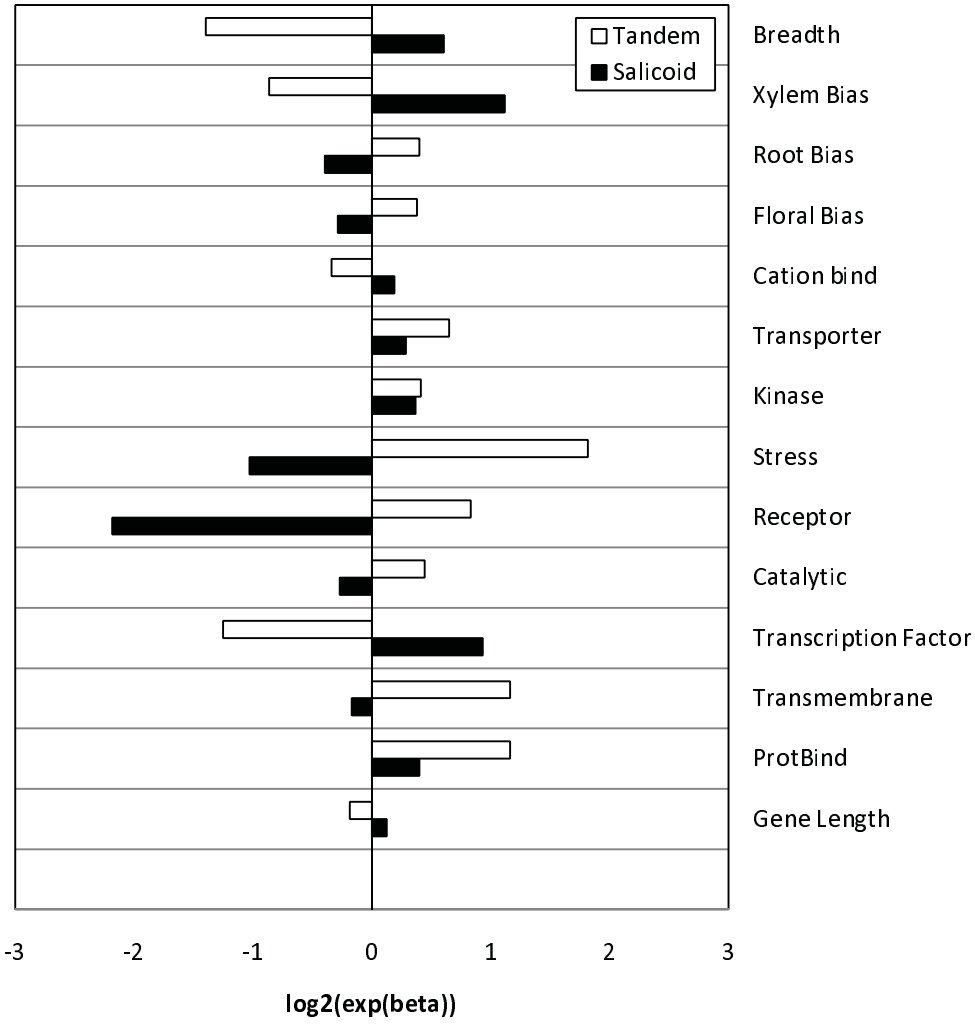
Warren AS, Anandakrishnan R, and Zhang L. 2010. Functional bias in molecular evolution rate of Arabidopsis thaliana. *BMC Evol. Biol.* **10**:125

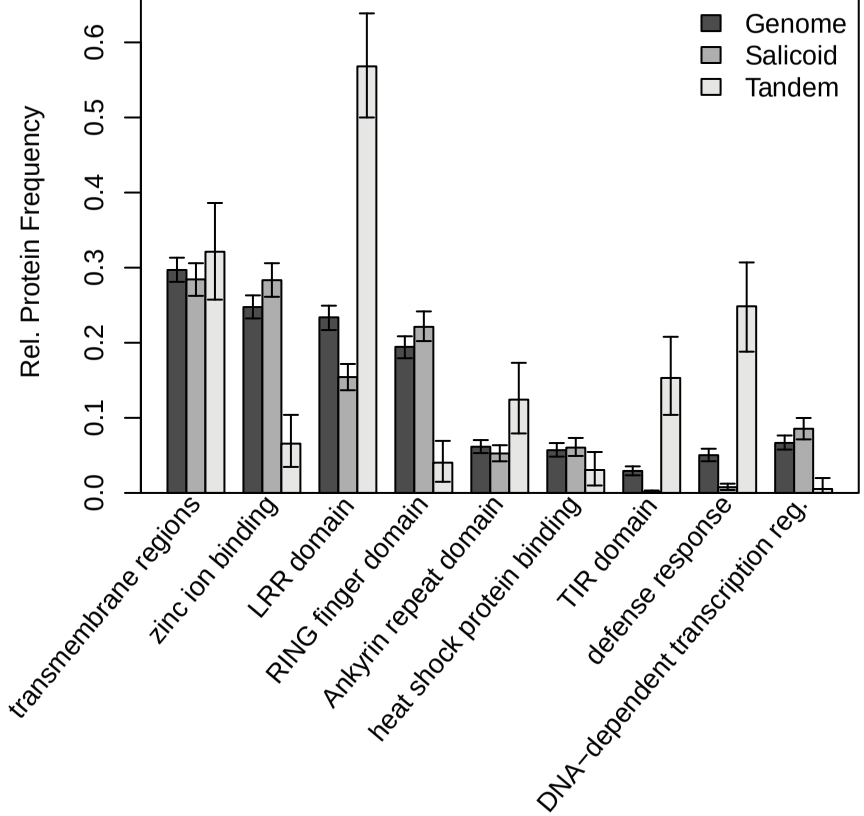
Woodhouse MR, Pedersen B, and Freeling M. 2010. Transposed Genes in Arabidopsis Are Often Associated with Flanking Repeats. *PLOS Genet.* **6**:e1000949

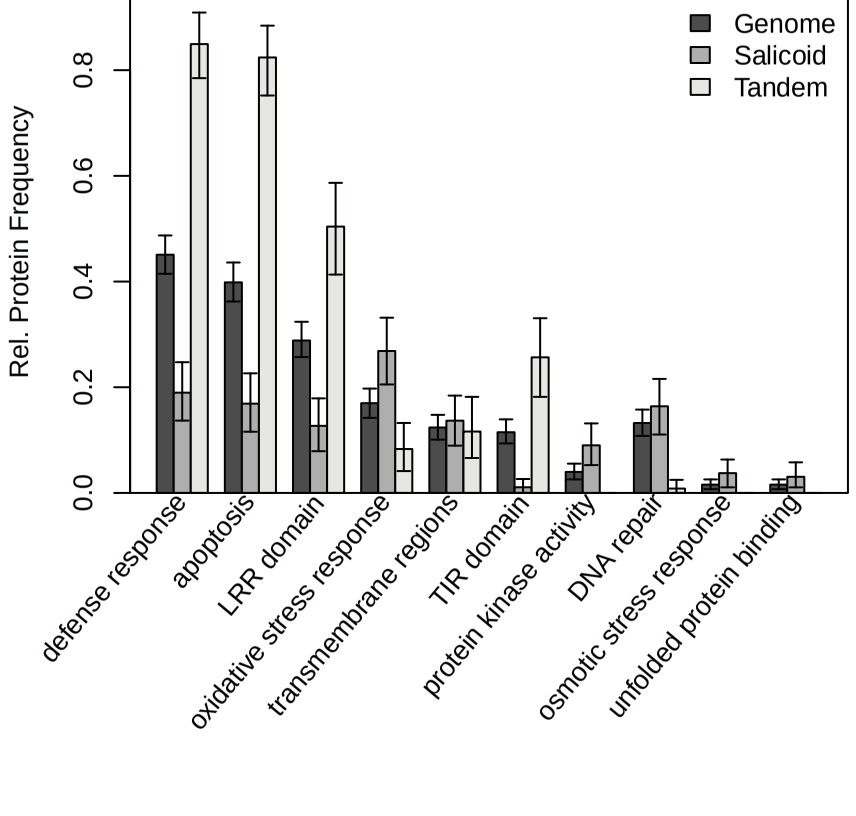
Yang ZH. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**:1586-1591

Zhang M, Wu YH, Lee MK, Liu YH, Rong Y, Santos TS, Wu C, Xie F, Nelson RL, and Zhang HB. 2010. Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors. *Nucleic Acids Res.* **38**:6513-6525

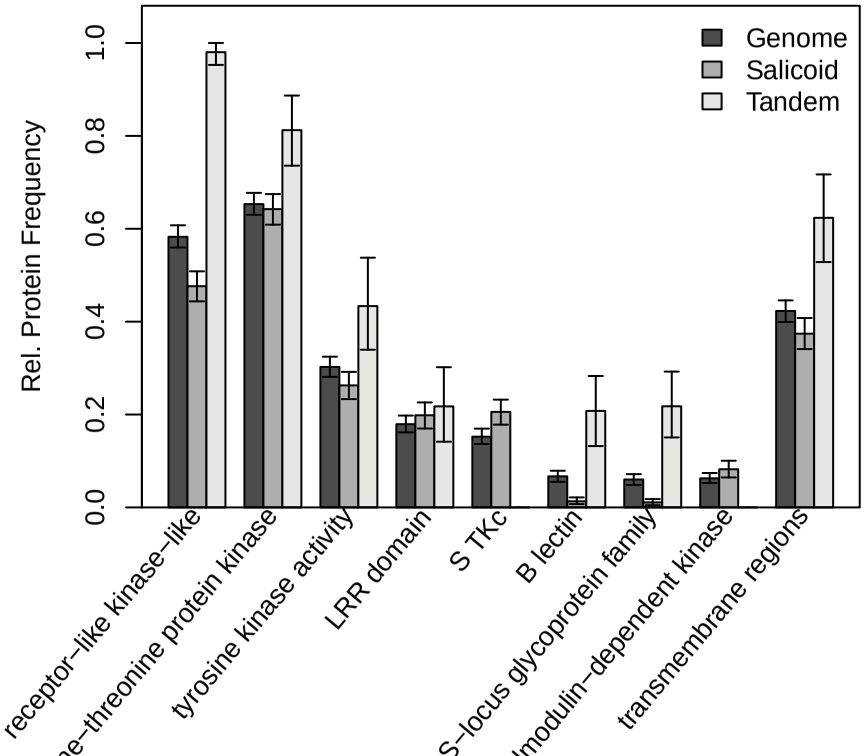
Zhang YJ, Wu YR, Liu YL, and Han B. 2005. Computational identification of 69 retroposons in Arabidopsis. *Plant Phys.* **138**:935-948



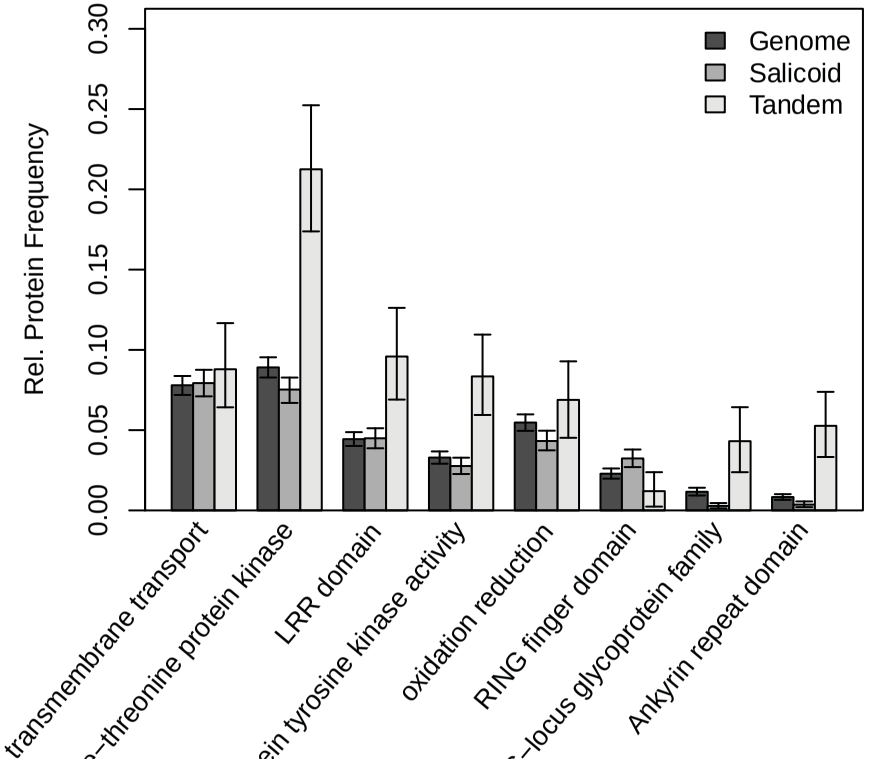

Rel. Protein Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Genome
Salicoid
Tandem
transmembrane regions
zinc ion binding
LRR domain
RING finder domain
Ankyrin repeat domain
heat shock protein binding
TIR domain
defense response
DNA−dependent transcription reg.

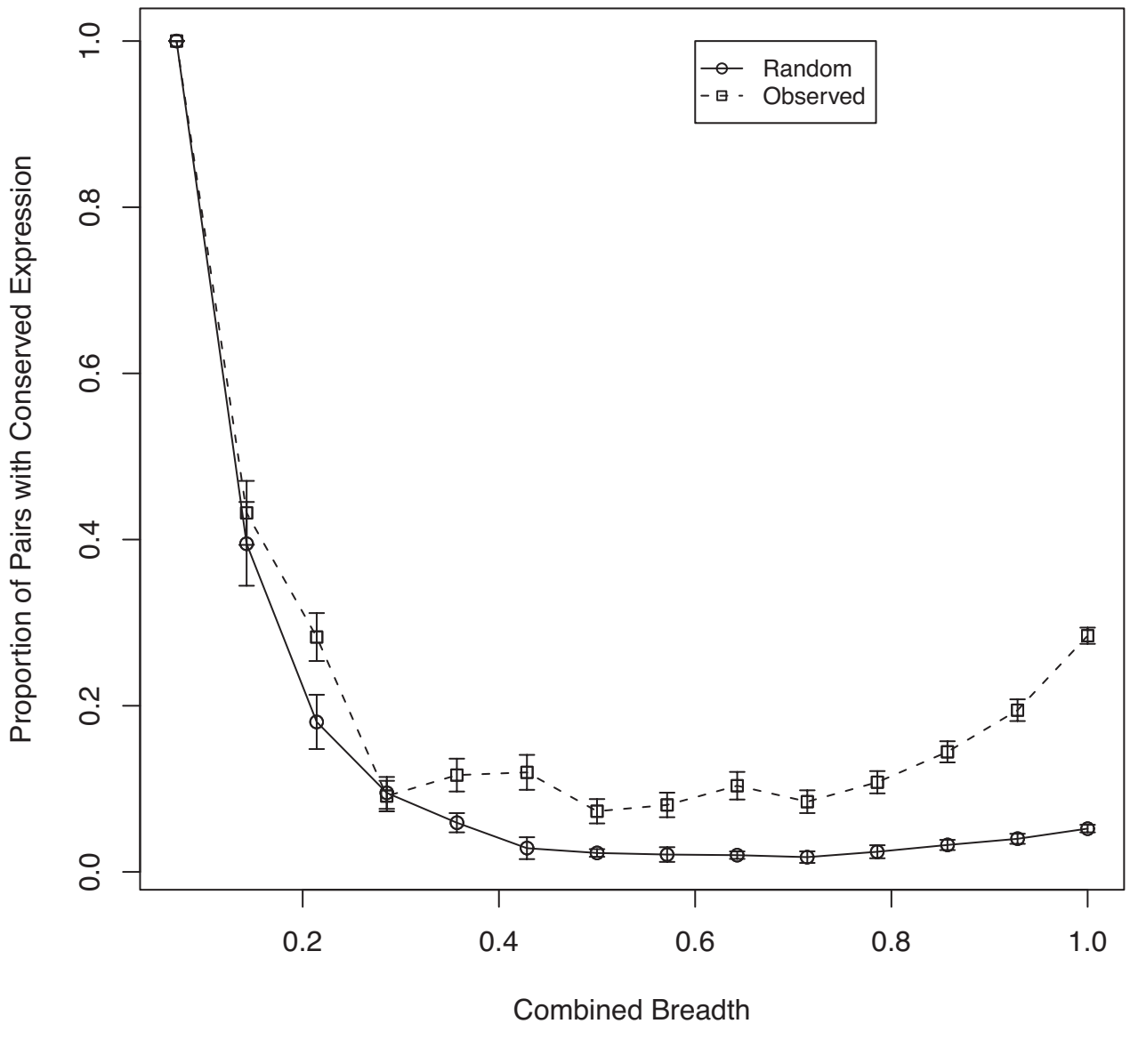

Rel. Protein Frequency
0.0
0.2
0.4
0.6
0.8
Genome
Salicoid
Tandem
defense response
apoptosis
LRR domain
oxidative stress response
transmembrane regions
TIR domain
protein kinase activity
DNA repair
osmotic stress response
unfolded protein binding

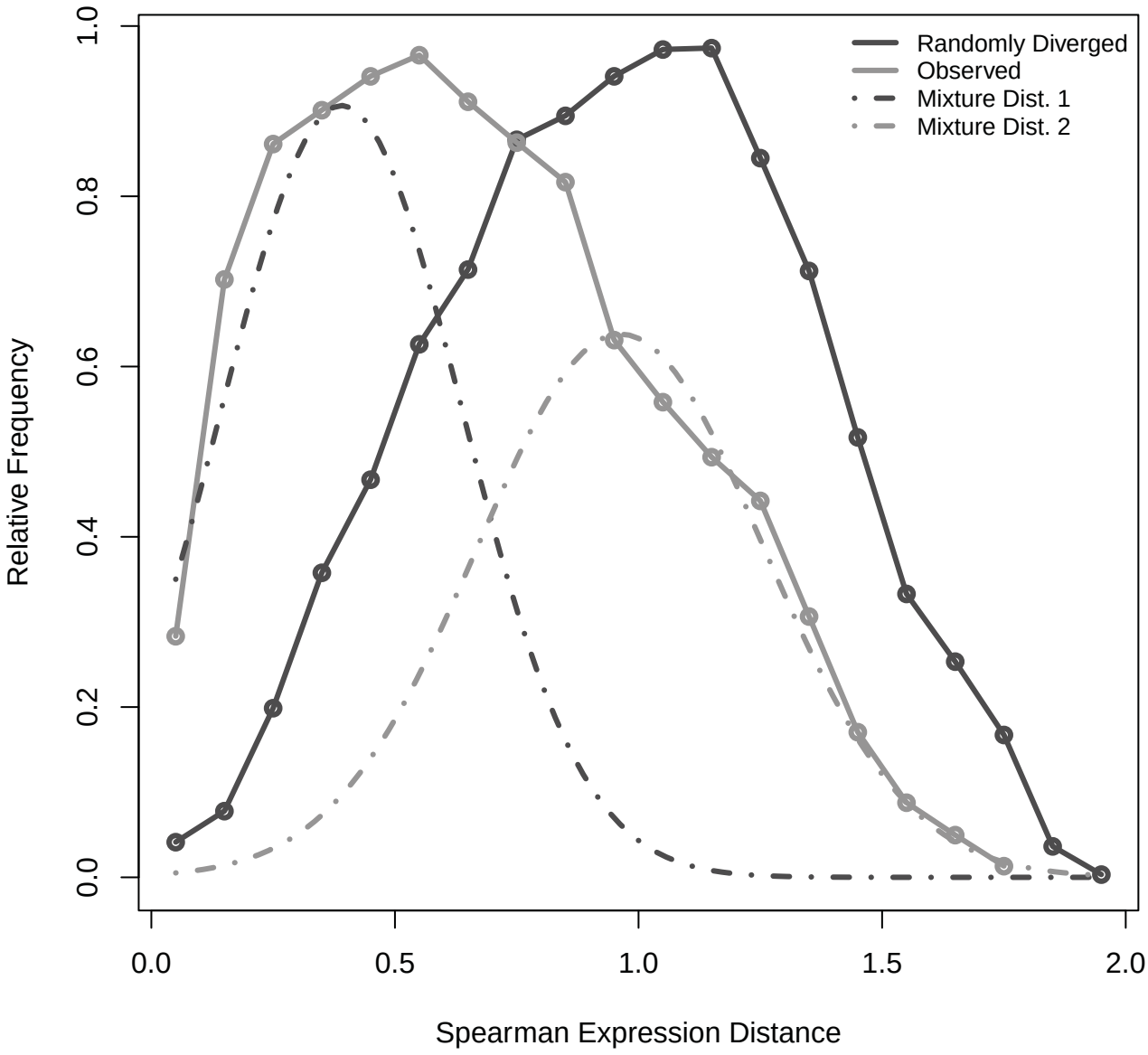
C.

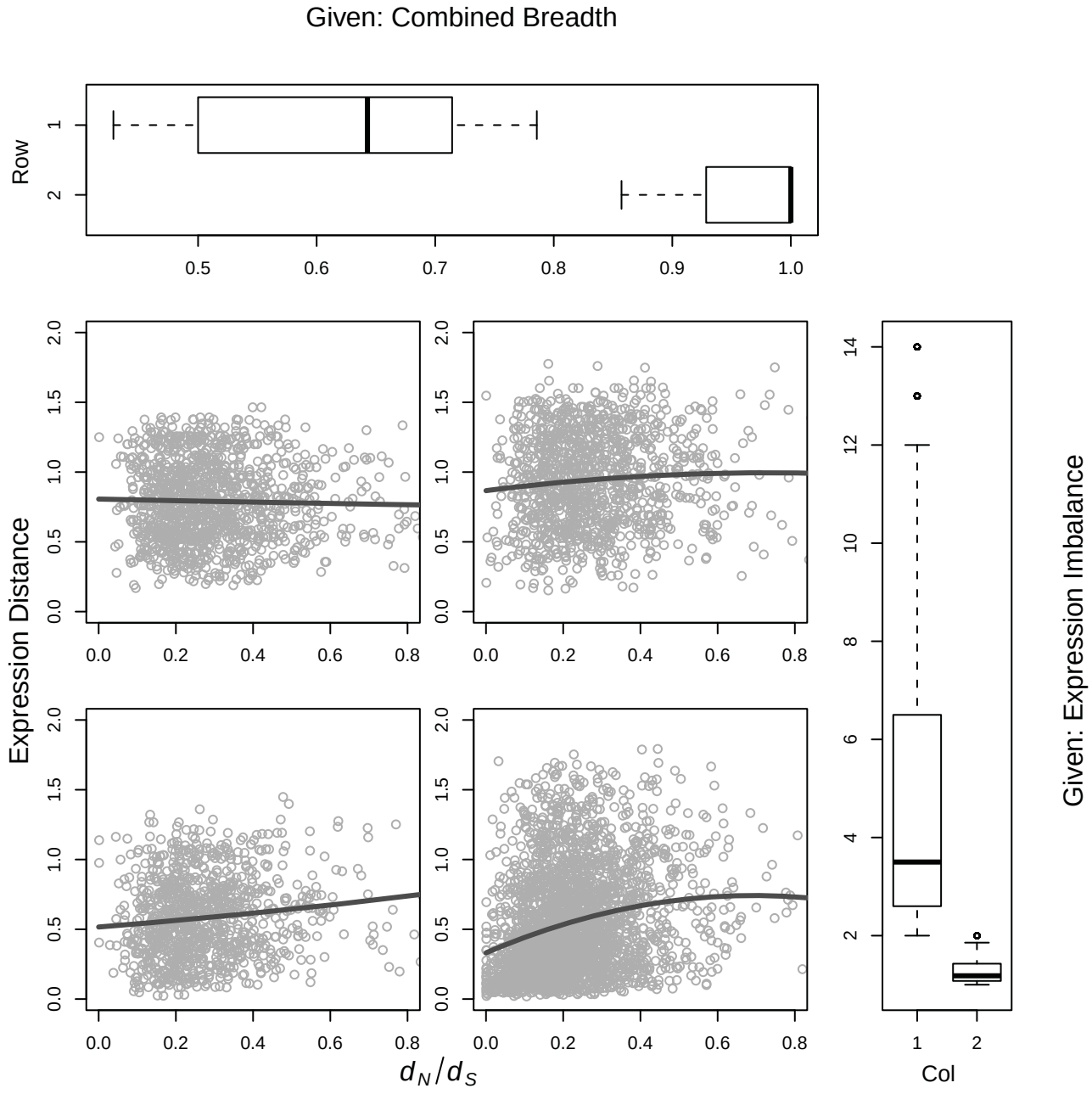

Rel. Protein Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Genome
Salicoid
Tandem
receptor−like kinase−like
tyrosine kinase activity
LRR domain
S TKc
B lectin
S−locus glycoprotein family
transmembrane regions

D.


Rel. Protein Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Genome
Salicoid
Tandem
transmembrane transport
LRR domain
oxidation reduction
RING finder domain
S−locus glycoprotein family
Ankyrin repeat domain




Randomly Diverged
Observed
Mixture Dist. 1
Mixture Dist. 2
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
Spearman Expression Distance


Given: Combined Breadth
Row
1
2
0.5
0.6
0.7
0.8
0.9
1.0
Expression Distance
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
d_N/d_S
Col
1
2
2
4
6
8
10
12
14
Given: Expression Imbalance