The evolution of the paleopolyploid genome in family Salicaceae

Eli Rodgers-Melnick
West Virginia University

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
https://researchrepository.wvu.edu/etd/3662

This Dissertation is brought to you for free and open access by The Research Repository @ WVU. It has been accepted for inclusion in Graduate Theses, Dissertations, and Problem Reports by an authorized administrator of The Research Repository @ WVU. For more information, please contact ian.harmon@mail.wvu.edu.
The evolution of the paleopolyploid genome in family Salicaceae

Eli Rodgers-Melnick

Dissertation submitted to the
Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
In
Biology

Stephen DiFazio, Ph.D., Chair
Vagner Benedito, Ph.D.
Jennifer Hawkins, Ph.D.
David Klinke, Ph.D.
Rita Rio, Ph.D.

Department of Biology

Morgantown, West Virginia
2013

Keywords: evolution; gene duplication; *Populus*; gene balance; indels; protein-protein interaction; networks; fractionation; paleopolyploid; polyploid

Copyright 2013 Eli Rodgers-Melnick
ABSTRACT

The evolution of the paleopolyploid genome in family Salicaceae

Eli Rodgers-Melnick

Whole genome duplication is a recurrent phenomenon throughout the history of flowering plants, with at least 1 occurring in the history of each angiosperm lineage. However, following duplication the genome typically undergoes a period of increased structural rearrangements and gene loss, resulting in a diploidized genome with large stretches of syntenic regions resulting from the duplication. This process, termed fractionation, often exhibits biased patterns of gene retention, which appear to correspond with the putative connectedness of genes within biological networks. In this thesis, I explore the extent to which the same biases appear in duplicate genes from the most recent whole genome duplication in *Populus trichocarpa*, how those biases relate to inferred network properties, and whether the ongoing process of nonfunctionalization within populations exhibits similar patterns to the historical trends. In Chapter 2, I compare the factors affecting retention of duplicates following both tandem and whole genome duplication. I show that these differ in a manner consistent with the action of a selective drive to maintain dosage balance. I also find that many duplicate pairs have more conserved expression patterns than expected by chance, consistent with the action of purifying selection for the maintenance of gene dosage. In Chapter 3, I use statistical learning to predict a protein-protein interaction network for *P. trichocarpa* and several model organisms. Using these predicted networks, I show that the probability of retention following several whole genome duplications in both *P. trichocarpa* and *Arabidopsis thaliana* is positively related to the relative abundance of duplicated neighbors and that the effect is enhanced with higher connectedness, consistent with a drive to maintain dosage balance between well-connected duplicates within biological networks. In Chapter 4, I use next generation sequencing to define small and large deletions within natural populations of *P. trichocarpa*. The patterns of enrichment among the deletions putatively causing nonfunctionalization events parallel the biases observed in chapters 2 and 3. This research therefore supports the importance of dosage balance during the fractionation process and illustrates how the observed biases continue to arise within natural populations.
Acknowledgements

The work contained herein would not have been possible without extensive support from friends, family, collaborators, and mentors. First and foremost, I would like to thank my advisor, Dr. Stephen DiFazio. Over the last 4 years he has provided unwavering support, guidance, and friendship. Dr. DiFazio’s incisive questions and unyielding scientific curiosity have pushed me to be a better scientist. His devotion and openness to his students’ interests have provided me with a model for mentorship that I will carry with me throughout my scientific career. I am continually impressed by his willingness and ability to follow me down novel avenues of research, and I am forever thankful for his commitment to my projects and scientific development. I was also fortunate to work with Dr. Gancho Slavov, whom I consider my second advisor. His sage advice regarding statistics and population genetics aided me greatly during my first year of doctoral research. I would also like to thank my former and current committee members, Dr. David Klinke, Dr. Vagner Benedito, Dr. Jennifer Hawkins, Dr. Rita Rio, and Dr. Olga Zhaxybayeva. Their guidance during my preliminary examination studies and their comments on my research proposals have broadened my scientific outlook and strengthened this manuscript.

I am thankful for the support and friendship of my fellow graduate students, especially the other members of the DiFazio lab: Rose Strickland-Constable, Hari Chetri, and Brahma Reddy Induri. I am particularly grateful for the friendship of Ryan Williams, whose support outside of the lab led me to this area of research with my sense of humor intact. Thank you to the research associates of the DiFazio lab: Dr. Luke Evans, Dr. Danielle Ellis, and Mindie Funke. Their questions, comments, and collaboration greatly supported my research. I would like to extend a special thank you to Dr. Evans, whose work ethic and positive attitude have given me an excellent role model as a young scientist. Thank you to the undergraduates who worked on various aspects of my projects and gave me valuable experience mentoring: Amanda Black, Joseph Tighman, George Magnone, and Sandy Simon. Their work ethic is a credit to undergraduates everywhere, and their output was indispensible to this dissertation.

I would like to sincerely thank my numerous collaborators on the chapters contained herein, whose data, leadership, and valuable insight made this work possible. I am especially grateful for the advice of Dr. Amy Brunner and Dr. Mark Culp for their support in the writing of chapters 2 and 3. I would also like to extend a special thank you to Dr. Culp and Dr. Philip Turk for sparking my interest in statistics. Dr. Turk’s encouragement initially gave me the courage to pursue a math-intensive career path, while Dr. Culp’s mentorship has provided me with a solid foundation for approaching myriad statistical problems over the course of my career. I would also like to offer a sincere thank you to Dr. Adam Halasz, who introduced me to the mathematical modeling of dynamical systems and whose encouragement inspired me to tackle
increasingly challenging problems. Lastly, I would like to thank my high school math teacher, Mrs. Ellen Flanigan, for showing me how to enjoy the language of mathematics for the first time.

Thank you to my family and friends outside the lab who encouraged me to pursue my passions and provided me with much-needed downtime. Thank you to my parents, who never dissuaded me from running around in the woods and carrying back all manners of flora and fauna. Thank you to my brother, who – in spite of having what is arguably the most fulfilling job in the world – never seems to stop bragging about his big brother the scientist. Finally, thank you to my wife, Sarah Kotcon. Sarah’s support originally gave me the courage to switch my field of study. Her encouragement and advice throughout my doctoral studies have been indispensible. She has toiled in the field, helped me during my studies, and – despite my best efforts – pulled me away from the computer long enough for me to eat. I am truly thankful that I get to spend every day with her.

This work was supported by the Office of Science (BER), U.S. Department of Energy, Grants No. DE-FG02-06ER64185 and DE-FG02-07ER64449, funding from the BioEnergy Science Center, a U.S. DOE Bioenergy Research Center (Office of Biological and Environmental Research in the DOE Office of Science), and through a United States Department of Agriculture National Institute of Food and Agriculture pre-doctoral fellowship. Additional research and travel funding was provided by the Office of the Dean of Eberly Arts and Sciences.
# TABLE OF CONTENTS

THE EVOLUTION OF THE PALEOPOLYPLOID GENOME IN FAMILY SALICACEAE ........................................................................................................... i

ABSTRACT .......................................................................................................................... ii

ACKNOWLEDGEMENTS ........................................................................................................ iii

TABLE OF CONTENTS .......................................................................................................... v

LIST OF FIGURES ............................................................................................................. vii

LIST OF TABLES ............................................................................................................... xii

CHAPTER I: GENERAL INTRODUCTION ........................................................................... 1
   References .......................................................................................................................... 17

CHAPTER II: CONTRASTING PATTERNS OF EVOLUTION FOLLOWING WHOLE GENOME VERSUS TANDEM DUPLICATION EVENTS IN POPULUS .................................................................................................................... 24
   Abstract ............................................................................................................................ 25
   Introduction ....................................................................................................................... 26
   Methods ............................................................................................................................ 30
   Results ............................................................................................................................... 38
   Discussion ........................................................................................................................ 53
   Acknowledgements ......................................................................................................... 61
   References ....................................................................................................................... 62
   Supplementary Materials ............................................................................................. 71

CHAPTER III: PREDICTING WHOLE GENOME PROTEIN INTERACTION NETWORKS FROM PRIMARY SEQUENCE DATA IN MODEL AND NON-MODEL ORGANISMS USING ENTS .................................................................................................................... 76
   Abstract ............................................................................................................................ 77
   Introduction ....................................................................................................................... 78
   Methods ............................................................................................................................ 81
   Results ............................................................................................................................... 91
   Discussion ........................................................................................................................ 111
   Acknowledgements ....................................................................................................... 117
   References ....................................................................................................................... 118
   Supplementary Materials ............................................................................................. 124
CHAPTER IV: DELETIONS REVEAL THE ONGOING PROCESS OF FRACTIONATION AND ENVIRONMENTAL ADAPTATION IN THE *POPULUS TRICHOCARPA* GENOME

Abstract .............................................................................................................................................. 139
Introduction ........................................................................................................................................ 140
Methods .............................................................................................................................................. 143
Results ................................................................................................................................................ 151
Discussion ......................................................................................................................................... 167
Acknowledgements ............................................................................................................................ 175
References ......................................................................................................................................... 176
Supplementary Materials ................................................................................................................... 183

CHAPTER V: GENERAL CONCLUSIONS ......................................................................................... 206
References ......................................................................................................................................... 212
LIST OF FIGURES

Figure 2.1. 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 .............................. 48

Figure 2.2. 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.................................................. 49

Figure 2.3. 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.......................................................... 50

Figure 2.4. 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 ($\mu_1$) = 0.3907, standard deviation ($\sigma_1$) = 0.2470, mixture proportion ($\pi_1$) = 0.5294. Parameters for the right distribution were $\mu_2$ = 0.9630, $\sigma_2$ = 0.2944, and $\pi_2$ = 0.4706.................................................................................................................................................. 51

Figure 2.5. Spearman expression distance plotted against $d_S/d_S$ 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 $d_S/d_S + (d_S/d_S)^2$ is shown in red for each subplot........................................................................................................................................... 52

Supplemental Figure 2.1. Boxplots for the $d_S$ ranges in Salicoid and tandem duplicate pairs ........................................................................................................................................ 71

Supplemental Figure 2.2. Mean $d_S/d_S$ for Salicoid and tandem duplicates grouped into bins of 0.2 $d_S$. Error bars indicate standard error of the mean .............................................................. 72

Supplemental Figure 2.3. Scatter plots of the replicates for the slides used in this analysis. Each plot contains a line of best fit based on least squares regression, along with the R-squared value ............................................................................................................................................... 73
Supplemental Figure 2.4. Plots of the replicates for two of the slides used in this analysis that had relatively poor correspondence between replicates. These were used for further analysis because they clustered with other floral tissues (Supplemental Figure 5) ................................. 74

Supplemental Figure 2.5. Hierarchical clustering based on the Euclidean distance for all replicates in the original microarray. VYL, SE, ApL, and Xpi slides were dropped for subsequent analyses ........................................................................................................................................ 75

Figure 3.1. The ROC curves for ENTS testing data on (A) *A. thaliana*, (B) *H. sapiens*, and (C) *S. cerevisiae* ........................................................................................................................................ 103

Figure 3.2. The number of positive predictions on sets of novel experimentally-determined interactions for ENTS and several alternative prediction techniques in (A) *A. thaliana*, (B) *S. cerevisiae*, (C) *H. sapiens*, and (D) *M. musculus*. ENTS results are shown in blue; the alternative classifiers’ results are shown in green; and the number of positive predictions shared between each pair of classifiers is shown in yellow ................................................................................................. 104

Figure 3.3. (A) The frequency of shared KEGG pathways and (B) the mean GO molecular function similarity scores for predicted ENTS interactions. Vertical lines at each confidence level show the total range for randomized networks. (C) Shared KEGG pathway frequency and (D) mean GO molecular function similarity for ENTS as compared to other predicted networks and the experimentally-verified network. The experimentally-verified networks are shown on the left for each organism, with a hatched bar ........................................................................................................................................ 105

Figure 3.4. Dots show the observed frequency of interactors with $\rho > 0.5$. Error bars indicate 95% confidence intervals based on 250 bootstrapped replicates. Grey bars indicate 95% confidence intervals from 250 randomized networks. The leftmost lane for each organism with the exception of *P. trichocarpa* is the data for the experimentally-verified network .................. 106

Figure 3.5. Each node represents a pathway in the PoplarCyc metabolic network, with lines connecting those with a significant number of interpathway interactions based on ENTS predictions. Red lines indicate a shared compound between pathways. The yellow nodes belong to a cluster significantly enriched for fatty acid biosynthesis, while the green and blue nodes belong to pathways significantly enriched for nucleotide/nucleoside biosynthesis and degradation, respectively ........................................................................................................................................ 107

Figure 3.6. (A) The cumulative distribution of literature similarity scores in randomized networks (green) and the observed human disease network constructed from ENTS PPI predictions (blue). (B) and (C) connected components in the human disease network. Line widths are drawn proportional to literature similarity scores ........................................................................................................................................ 108

Figure 3.7. The relationship between degree centrality, the fraction of predicted neighbors retained following WGD, and the probability of whether a gene in the predicted network is
retained following the WGD for the (A) *P. trichocarpa* Salicoid duplication, (B) the *A. thaliana* α duplication, and (C) the *A. thaliana* β and γ duplications ................................................................. 109

**Supplemental Figure 3.1.** Kernel density plots of the highest domain odds scores for interacting and non-interacting protein pairs in *A. thaliana* (a), *H. sapiens* (b), and *S. cerevisiae* (c) ...... 130

**Supplemental Figure 3.2.** The total number of interactions predicted by ENTS compared to other classifiers. ENTS results are shown as horizontal lines, with colors corresponding to confidence thresholds as follows: 0.55 (yellow), 0.65 (light orange, 0.75 (dark orange), 0.85 (red) ............................................................................................................................ 131

**Supplemental Figure 3.3.** The total number of genes with predicted interactions from ENTS compared to other classifiers ................................................................................................................................. 132

**Supplemental Figure 3.4.** The mean GO biological process and (B) cellular component similarity scores for predicted ENTS interactions. Vertical lines at each confidence level show the total range for randomized networks. (D) Shared GO biological process and (C) cellular component similarity for ENTS as compared to other predicted networks ............................. 133

**Supplemental Figure 3.5.** Variable importance for the first 30 most important variables in the *A. thaliana*-trained ENTS classifier. Importance was measured as the mean decrease in OOB classification accuracy during training following permutation of the given variable .............. 134

**Supplemental Figure 3.6.** Variable importance for the first 30 most important variables in the *S. cerevisiae*-trained ENTS classifier .......................................................................................... 135

**Supplemental Figure 3.7.** Variable importance for the first 30 most important variables in the *H. sapiens*-trained ENTS classifier .......................................................................................... 136

**Figure 4.1.** Cumulative amount of sequence space within the large deletions, small deletions, and genome composed of (A) repeat elements, (B) genes and conserved noncoding sequence, and (C) genes with and without retained Salicoid duplicates. (D) SNP density within the region adjacent to small deletions .................................................................................................................. 160

**Figure 4.2.** Size and MAF distributions of small deletions within CDS regions. (A) Observed (black) and expected (grey) counts of CDS small deletions. (B) The range of differences between the observed median MAF and the range of expected MAFs given the background distribution for that deletion size. (C) Observed and expected MAFs for small and large deletions overlapping genes with and without retained Salicoid duplicates. Error bars indicate 95% confidence intervals. (*** p < 0.001, ** p < 0.01, * p < 0.05; Two-tailed) ........................................ 161

**Figure 4.3.** Differences in median tissue specificity between Salicoid duplicates in which one member of the pair is affected by (A) a small frameshift deletion or (B) a large deletion. Blue
dots indicate the observed median difference between the tissue specificity values, and grey bars indicate 95% confidence intervals for no true difference in specificity..........................162

Figure 4.4. Enrichment of plant GO-SLIM categories among genes with deletions affecting their CDS. Highly positive Z-values (red) indicate an enrichment of the category, while highly negative Z-values (blue) indicate an underrepresentation of the category ............................163

Figure 4.5. (A) Bootstrapped median degrees for Salicoid catalytic genes in the poplarCyc metabolic network. (B) Observed (black) and expected (grey) numbers of Salicoid duplicate pairs among nucleic acid binding genes containing small non-frameshift deletions. (C) Positions of small non-frameshift deletions within a group of myb-like transcription factors. Horizontal lines indicate sequences within a multiple sequence alignment, while red triangles indicate the positions of the deletions. Sequence conservation at each position is given by the vertical lines at the top, and the DNA-binding myb repeats are highlighted in blue ........................................164

Figure 4.6. (A) QQ-plot and (B) Manhattan plot of deletion associations with the 4th environmental PC. Dotted horizontal lines indicate $p = 0.05$, as assessed by 10,000 permutations. ............................................................................................................................165

Supplemental Figure 4.1. Schematic of PCR validation assay for large deletions and an example assay gel ......................................................................................................................185

Supplemental Figure 4.2. The first two principal components acquired using only deletion polymorphisms and colored by the population where they were originally collected ............186

Supplemental Figure 4.3. The first two principal components acquired using only SNPs and colored by the population where they were originally collected .............................................187

Supplemental Figure 4.4. Kernel density estimates of the fixation index for SNPs, small deletions, and small insertions discovered using mpileup. Positive values indicate an excess of homozygosity relative to that expected under Hardy-Weinberg Equilibrium .......................188

Supplemental Figure 4.5. Density of small deletions within and around recombination hotspots. The green dot gives the mean density of small deletions in deletions per base pair for 606 recombination hotspots. Blue dots give the mean density for 2500bp non-overlapping bins extending away from the boundaries of the hotspot region. The highlighted blue range indicates a 95% CI for the expected mean indel density in recombination hotspots based on permutations throughout the genome ..............................................................................................................189

Supplemental Figure 4.6. Gene expression for genes with CDS impacted by deletions. (A) Mean expression values for genes impacted by deletions (black) vs. the expected range of mean values (grey). (B) Mean Pearson coexpression between Salicoid duplicates impacted by deletions (black) vs. expected. Error bars indicate 95% confidence intervals. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; Two-tailed) ..................................................................................................................190
Supplemental Figure 4.7. Observed (black) and expected values for Salicoid frameshift deletions .................................................................191

Supplemental Figure 4.8. Observed (black) and expected values for Salicoid non-frameshift deletions ........................................................................................................192

Supplemental Figure 4.9. Observed (black) and expected values for Salicoid large deletions ........................................................................................................193

Supplemental Figure 4.10. Observed (black) and expected values for non-Salicoid frameshift deletions .....................................................................................194

Supplemental Figure 4.11. Observed (black) and expected values for non-Salicoid non-frameshift deletions .....................................................................................195

Supplemental Figure 4.12. Observed (black) and expected values for non-Salicoid large deletions .................................................................................................196

Supplemental Figure 4.13. Positions of small non-frameshift deletions within the largest group of bHLH transcription factors. Horizontal lines indicate sequences within a multiple sequence alignment, while red triangles indicate the positions of the deletions. Sequence conservation at each position is given by the vertical lines at the top, and the DNA binding domain is highlighted in blue ....................................................................................................................197

Supplemental Figure 4.14. Positions of small frameshift deletions in the largest group of WRKY transcription factors. WRKY DNA binding domains are highlighted in blue ............198

Supplemental Figure 4.15. Scree plot of the principal components of environmental variation ..................................................................................................................199

Supplemental Figure 4.16. Maps of the top 4 PCs of environmental variation. Red hues indicate higher values. Black dots signify sampling locations for the trees used in the association analysis. ........................................................................................................................200

Supplemental Figure 4.17. The number of unique genes implicated in associations with the first 4 PCs of environmental variation when performed on SNPs only and deletions only. Genes were deemed implicated if significantly associated markers resided within the genic regions or within the intergenic region flanking them .................................................................................................................201

Supplemental Figure 4.18. Deletion association results for the environmental PC1. The horizontal lines indicate the 95% significance level based on permutation ..............................................202

Supplemental Figure 4.19. Deletion association results for the environmental PC2 ............203

Supplemental Figure 4.20. Deletion association results for the environmental PC3 ..........204
LIST OF TABLES

Table 2.1. Significant predictors for retention of genes in Salicoid and tandem duplicate pairs ................................................................. 44

Table 2.2. Mean, median, and standard error for the gene length, expression breadth, and \(d_S/d_S\) (of pairs) within the whole genome, among Salicoid duplicates, and among TDs. P-values for gene length and breadth are based on van der Waerden tests for independence of gene characteristic and duplicate type................................................................................................................................. 45

Table 2.3. Proportion of genes annotated with a given functional category in each of the 2 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................................................................................................................................. 46

Table 2.4. Regression table for the relationship between expression distance and \(d_S/d_S\), combined breadth, and expression imbalance. The estimate refers to the linear regression coefficient for each predictor variable................................................................................................................................. 47

Table 3.1. Protein-protein interactions predicted using ENTS at several confidence cutoffs. The number of genes involved in the predicted interactions is shown in parentheses ......................... 100

Table 3.2. Classes of metabolic pathway that were significantly enriched in clusters at a 0.05 family-wide type I error rate. The number of pathways present in each cluster is given in parentheses in the first column, while the number of the given class of pathway within each cluster is given by parentheses in the second column ................................................................................................................................. 101

Table 3.3. Estimated coefficients and significance for the logistic regressions on WGD retention in \(P. trichocarpa\) and \(A. thaliana\) using network properties degree centrality (DC) and duplicated neighbors (DN) from the ENTS 0.65 confidence networks. P-values are based on confidence intervals generated from fitting the model on resampled data 10,000 times ................................. 102

Supplemental Table 3.1. Predictors used for the ENTS classifier. Importance measures indicate the mean decrease in accuracy in the out-of-bag training data upon permutation of that variable. ................................................................................................................................................................................................. 123

Supplemental Table 3.2. 95% confidence intervals for coefficients in the logistic regressions on WGD retention in \(P. trichocarpa\) and \(A. thaliana\). Models were fit using either all genes with at least one interaction in the ENTS predicted network or with the subset with at least 10 neighbors. Confidence intervals were based on non-parametric bootstrapping with 10,000 random resamplings ................................................................................................................................................................................................. 129

Table 4.1. Linear model explaining small deletion frequency in 250kb windows across the genome ................................................................................................................................................................................................................................................................................................. 158
**Table 4.2.** 95% confidence intervals for the effects of Salicoid duplicate status and deletion type on expression level. Duplicate status effect is relative to non-Salicoid duplicates, while deletion type effects are relative to frameshift deletions. Confidence intervals were generated by 10,000 bootstrapped samples ................................................................. 159

**Supplemental Table 4.1.** Results of PCR validation assays. Column headings titled as XX/XX indicate PCR result/prediction. The insertion and deletion alleles are represented by the letters “I” and “D”, respectively ................................................................. 182

**Supplemental Table 4.2.** Spearman correlations (below diagonal) and p-values (above diagonal) between variables in 250kb windows across the genome. Small deletions, large deletions, SNPs, gene content, and CpG islands indicate the counts for those elements, while GC content and N content represent the percentage of GC and N nucleotides within the window ................. 183

**Supplemental Table 4.3.** Pearson correlations of environmental PCs with individual environmental variables at the sampling locations. The most significant positive and negative correlations for each PC are highlighted with bold boxes ................................................................. 184
CHAPTER I

GENERAL INTRODUCTION
Whole Genome Duplication

Whole genome duplications (WGDs) have provided a recurrent source of evolutionary novelty throughout the history of angiosperms and other seed plants (Jiao et al. 2011). At the population level, WGDs can reproductively isolate neopolyploids from their non-polyploid source populations, thereby providing a simple and plausible means for sympatric speciation to occur (Otto and Whitton, 2000; Mable, 2004). The doubling of the genome is also associated with several immediate physiological phenomena that may contribute to niche adaptation among neopolyploids. These include increases in cell volume and decreases in cellular division rates (Gregory and Hebert, 1999), which may have substantial impacts on metabolic rates, organ growth, and plant-animal interactions (Lou and Baldwin, 2003; Comai, 2005).

The long-term success of post-WGD lineages may depend on a number of factors, many of which are still under investigation. Among the most important may be the mechanism of polyploidy formation, which involves a union of two diploid genomes from either separate species or the same species. In the first case, the offspring is known as an allopolyploid, while in the second it is called an autopolyploid. Historically, autopolyploids were thought to be very rare (Soltis et al. 2003). However, over the past twenty years, the estimate of autopolyploid frequency has increased. Indeed the spontaneous formation of autopolyploids is on par with the gene mutation rate, with a frequency in angiosperms of approximately 1 in 100,000 (Ramsey and Schemske, 1998). Despite this, allopolyploids are still thought to be more common (Soltis et al. 2003). The source of this discrepancy is debated. However, the relative infrequency of autopolyploids may result in part from difficulties that arise during the segregation of homologous chromosomes in anaphase I. The presence of more than 2 highly similar homologues within the autopolyploid may cause multivalents to arise following the formation of
chiasma during metaphase I (Cifuentes et al. 2010). The presence of multivalents may then increase the probability of non-balanced segregation during anaphase, which ultimately promotes the production of aneuploid gametes. Indeed, studies of autotetraploid maize and *Arabidopsis thaliana* demonstrate high rates of aneuploidy among the autopolyploid progeny (Doyle, 1986; Wei and Zhang, 2010). These aneuploidies can lead to deleterious dosage imbalances among functionally related components of biological networks and thereby limit the productivity of autopolyploids (Birchler and Veitia, 2012).

The relative success of allopolyploids may also be partially attributed to the effects of hybridization that occur independently from changes in ploidy. The fixed union of two divergent genomes that occurs within allopolyploids allows the neopolyploid to preserve heterozygosity at loci with interspecies polymorphisms even in the presence of severe inbreeding. Such fixed differences may induce a variety of changes in gene regulation, including those that result in transgressive expression relative to the parents (Flagel and Wendel, 2010; MacArthur and Tyler-Smith, 2010). Such patterns are postulated to provide a molecular basis for heterosis, the phenomenon whereby the progeny of hybrid crosses outperform either of the parental lines (Zhang et al. 2008).

**Fractionation**

Following whole genome duplication, the neopolyploid generally undergoes a period of increased genomic instability. This process, termed fractionation, ultimately returns the polyploid genome to a state of similarity with the ancestral diploid genome. In synthetic and natural allopolyploids, this process is characterized by decreases in the formation of multivalents during meiosis, deletions of repetitive and non-repetitive sequences, and large-scale structural rearrangements (Ozkan and Feldman, 2009; Mandakova et al. 2010; Woodhouse et al. 2010).
Classically, many thought most duplicated genes would be eliminated during fractionation soon after the WGD occurred, as they were assumed redundant apart from their ability to mask against deleterious mutations (Haldane, 1933). However, Susumu Ohno’s pioneering tome, Evolution by gene duplication, stressed the importance of duplicates as a source of evolutionary novelty, proposing two rounds of whole genome duplication at the stem of the vertebrate lineage based on the presence of 4 HOX gene clusters and several apparent duplications within the human genome (Ohno, 1970). The emergence of nucleotide sequencing technologies led to a resurgence of interest in Ohno’s work, as hundreds to thousands of functional gene duplicates were found within diverse organisms without recent whole genome duplications (Sidow, 1996; Wolfe and Shields, 1997; Initiative, 2000).

The discovery of high duplicate retention rates in paleopolyploid genomes spurred an expansion of theoretical research to explain how duplicate preservation could occur in the absence of a strong selective advantage for mutational masking or a high frequency of beneficial mutations (Lynch et al. 2001). Furthermore, any plausible theory would need to provide a rationale for the biased nature of the retained duplicate gene content within paleopolyploid genomes. This bias primarily refers to the overrepresentation of genes involved in conserved intracellular pathways, particularly core transcriptional and signal transduction networks. Such genes are found to be preferentially retained following WGDs in plants (Thomas et al. 2006; Buggs et al. 2012), fungi (Davis and Petrov, 2005), and animals (Hughes and Liberles, 2008), providing strong evidence for the action of selection on certain functional classes of duplicates during fractionation.

The Gene Balance Hypothesis
The biased retention of duplicates involved in core cellular networks following WGD suggests that the interactions between the proteins encoded by retained genes might constrain their copy numbers following duplication. Such an idea is further motivated by experimental work showing a positive association between a protein’s involvement in macromolecular complexes and the fitness defects accompanying manipulation of the protein’s dosage (Papp et al. 2003). These observations motivated the development of the gene balance hypothesis, which proposes that a selective constraint for the maintenance of stoichiometric balance between interacting proteins drives biased patterns of retention and loss following gene duplication (Birchler and Veitia, 2007; Birchler and Veitia, 2010).

The origins of the gene balance hypothesis begin prior to the genomics era, when multiple studies observed stark differences between the phenotypic consequences of unbalanced versus balanced shifts in ploidy, with the latter generally found to be less severe (Blakeslee et al. 1920; Birchler, 1979). These results generalized to studies of gene expression, wherein an increase in the dosage of a single gene would cause a concomitant increase or, more commonly, a decrease in its expression relative to the diploid. Conversely, a balanced dosage shift between the gene and its regulator would maintain the diploid expression level (Birchler et al. 2001). When extrapolated to evolution after gene duplication, these results imply that shifts in gene dosage mediated by changes in copy number due to either duplication or deletion may carry substantial fitness costs, especially when genes exhibit high dosage sensitivity due to central roles within biological networks.

The gene balance hypothesis predicts that small-scale (i.e. tandem and segmental duplications) and large-scale (i.e. WGDs) duplications should have contrasting biases in the types of functional classes retained (Veitia et al. 2008). This supposition is justified by the
differing dosage consequences of each type of duplication. Through the duplication of every gene in the progenitor genome(s), WGDs also identically increase the dosage of all genes and thereby maintain the ancestral stoichiometry of biological networks. The return of a dosage-sensitive gene to the diploid copy number during the fractionation process could then cause deleterious stoichiometric imbalances. Therefore, preservation of the duplicate would occur through purifying selection. By contrast, the duplication of a dosage-sensitive gene in isolation through a tandem or segmental duplication would also result in a deleterious stoichiometric imbalance, leading to selection against small scale duplications of the same gene classes well-retained following WGDs.

Dosage sensitivities may arise in a variety of network contexts. The best-studied context consists of physically interacting complexes, wherein either the duplication or the deletion of a subunit with more than 1 interaction can significantly lower the efficiency of full complex assembly (Veitia, 2004). Dosage sensitivity may also occur within transcriptional networks, signal transduction networks, or metabolic networks even in the absence of direct physical interactions with other proteins. For instance, within the mitogen activate protein kinase (MAPK) cascade, an increase in the dosage of either MAPKK or MAPK-phosphatase in isolation will significantly alter the shape of the sigmoidal response curve that ordinarily triggers transcriptional activation, while an increase in the expression of both will maintain the wild-type response (Veitia, 2005). Likewise, the proper functioning of transcriptional networks depends on the relative abundance of transcriptional activators and repressors. A non-balanced alteration of transcription factor dosage may therefore induce macro-scale changes in the long-term behavior of the network, such as converting an oscillatory circuit into a system with a stable steady-state (Veitia, 2005).
Subfunctionalization

Although the gene balance hypothesis may account for the biased-retention of dosage-sensitive duplicates following whole genome duplication, it does not provide a possible mechanism for retention of non-dosage-sensitive genes or an explanation for functional divergence that may occur post-duplication. Subfunctionalization, as proposed by Allan Force et al. in 1999, provides a model for both duplicate retention and divergence that is largely dependent upon the action of degenerative mutations and genetic drift (Force et al. 1999). Therefore, in contrast to gene balance or neofunctionalization, it allows for preservation of gene duplicates in the absence of direct selection for multiple copies.

Similar to the gene balance hypothesis, subfunctionalization as a mechanism for duplicate preservation was motivated by the higher than expected rates of duplicate retention within paleopolyploid genomes (Force et al. 1999). The hypothesis, more specifically called duplication-degeneration-complementation (DDC) model, proposes that if duplicate genes carry multiple, independently mutable subfunctions, the loss of subfunction A in duplicate 1 combined with the loss of subfunction B in duplicate 2 will lead to purifying selection for functional copies of both duplicates in order to maintain the ancestral suite of subfunctions. The definition of subfunctions is deliberately generalized, as they may refer to temporal or tissue-specific cis regulatory elements or to biochemical properties of the polypeptide (Force et al. 1999). However, the regulatory elements are generally thought to have a greater potential for independent mutability (Carroll, 2008).

This mechanics of the DDC process give the subfunctionalization hypothesis two desirable theoretical properties. First, in contrast to neofunctionalization, which requires the occurrence of relatively infrequent beneficial mutations prior to nonfunctionalization, subfunctionalization only
requires degenerative mutations to occur (Lynch and Walsh, 1998; Force et al. 1999). Second, the spread of the duplicate-preserving degenerative mutations through the population is mediated by the action of genetic drift rather than positive selection, as in the case of neofunctionalization. Therefore, subfunctionalization would be most effective in species with relative small effective population sizes, wherein fixation by drift can occur most quickly (Force et al. 1999; Lynch and Force, 2000).

Although Force and Lynch did not directly describe a mechanism for the biased retention of certain classes of duplicates through subfunctionalization, the DDC process does imply that those genes with higher regulatory complexity should have more independently mutable subfunctions and, therefore, a higher probability of duplicate retention. Moreover, there may be a trade-off between the probability of subfunctionalization and dosage constraints, as the degeneration of some duplicate subfunctions may have fitness costs if sensitive stoichiometric balances are altered (Force et al. 1999).

**Empirical studies of Network Context and Duplication**

While multiple studies of paleopolyploid genomes have demonstrated a positive association between post-WGD retention and categories of genes generally described as having higher connectivity (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al. 2005), empirical studies of experimentally-defined biological networks are less common, especially in plants where functional data is sparse. In the yeast *Saccharomyces cerevisiae*, which does have well-characterized protein-protein interaction, genetic, and metabolic networks, retained WGD duplicates are more frequently members of protein complexes than small scale duplicates (Hakes et al. 2007). They also tend to have greater levels of protein under-wrapping than small-scale duplicates, indicating that WGD duplicates require more physical interactions on average in
order to maintain protein stability (Liang et al. 2008). Moreover, yeast WGD duplicates share physical interactions and retain the same functional relationships between interactors more often than small-scale duplicates (Guan et al. 2007). Although these observations seem to support the importance of gene balance in maintaining stoichiometric balance among subunits within protein complexes, surprisingly WGD duplicates are found to be less essential than either singletons or small-scale duplicates, as assessed through high-throughput synthetic lethality experiments. However, the deletion of both members of a duplicate pair has a higher level of synthetic lethality in WGDs and small-scale duplicates (Hakes et al. 2007), indicating that while retained WGD duplicates may functionally compensate for each other, their combined functions are more likely to be essential. WGD duplicates are also found to exhibit lower coexpression than small-scale duplicates (Guan et al. 2007), indicating that if the WGDs were retained due to dosage constraints, substantial expressional divergence took place following the WGD.

Studies of the relationship between duplicability and network properties have also been carried out in plants, primarily within the model species Arabidopsis thaliana. An investigation of the A. thaliana metabolic network found that WGD duplicates retained from the most recent α-WGD are significantly clustered together, unlike those from the more ancient β-WGD. However, duplicates from the β-WGD have significantly higher metabolic flux than expected by chance (Bekaert et al. 2011). These results suggest that dosage balance may have constrained the evolution of the A. thaliana metabolic network following the most recent WGD by maintaining stoichiometric balance between reactions within pathways. However, the non-clustered retention of high flux reactions following the β-WGD is more indicative of selection for increased absolute dosage than the relative dosage regime advocated by the gene balance hypothesis. Nonetheless, these differences may reflect selection for relative dosage following each WGD, followed by
subsequent functional divergence and eventual preferential loss of duplicates mediating low-flux reactions (Bekaert et al. 2011).

Functional divergence of WGD duplicates in plants is also supported by a high-throughput study of the A. thaliana protein-protein interaction network. As with yeast, duplicate pairs from WGDs share significantly more interactors than pairs from small-scale duplications, and pairs with high coexpression share more interactors than pairs with low coexpression (Braun et al. 2011). However, WGD duplicates also exhibit considerable functional divergence, averaging 45% as the fraction of shared interactors between duplicates. Therefore, even if initial retention is mediated by dosage balance constraint, subsequent evolution of the interaction network conforms to the expectations of a subfunctionalization or neofunctionalization model.

Prediction of Protein-Protein Interactions

Protein-protein interaction (PPI) networks have a variety of uses for both functional and evolutionary studies. Following the elucidation of network structure, the relationships characterized by the PPI network topology may be harnessed to infer the molecular functions of individual proteins (Sharan et al. 2007), discover higher-order interactions between entire biochemical pathways (Dotan-Cohen et al. 2009), and evaluate the potential physiological consequences of perturbations to the network (Guney and Oliva, 2012). Furthermore, as illustrated by the characteristics of retained duplicates, PPI networks may provide considerable insight into the evolutionary process following duplication. However, comprehensive studies of empirical PPI networks are still largely confined to key model organisms: Homo sapiens (Stelzl et al. 2005), S. cerevisiae (Yu et al. 2008), and D. melanogaster (Guruharsha et al. 2011). Recently, the first high-throughput study of A. thaliana was conducted, which added over 5,000 interactions to its experimentally-determined interactome. Despite this advance in knowledge of
the plant interactome, the available interactions are still only thought to cover approximately 2% of the estimated ~300,000 total PPIs (Braun et al. 2011).

The usefulness of PPIs has led to the development of variegated techniques for the expansion of the interactome, especially within plants where high-throughput studies are rare. The simplest of these uses an interolog approach, wherein interactions are inferred between proteins if they have conserved orthologs known to interact within another species. While this approach does produce predicted interactions that are enriched in genes with high coexpression and semantic similarity within the gene ontology, such predictions are restricted to those genes most conserved among eukaryotes when the phylogenetic distances between orthologous networks are large. Therefore, the sizes of the interolog interactomes for *A. thaliana* based on *H. sapiens, C. elegans, D. melanogaster, and S. cerevisiae* interactions are relatively small (Geisler-Lee et al. 2007; De Bodt et al. 2009). However, the interolog approach may work well when a relatively closely related species has high interactome coverage (Yellaboina et al. 2008).

A number of researchers have attempted to circumvent the conservative nature of the interolog approach by combining functional data associated with physical interaction and using it to train classifiers that can then associate each potential interaction with a confidence score. In *A. thaliana*, both naïve Bayes and support vector machines (SVMs) have been used to associate functional predictors such as coexpression, subcellular localization, and experimentally-derived GO annotation with sequence-derived features such as domain composition and co-evolution in order to predict other interactions within that species (Cui et al. 2008; Xu et al. 2010). While these methods have produced larger predicted interactomes, their input features are often tailored to the model organism and require functional data that is often not available for non-model species. Therefore, these methods do not generalize well, and ready-to-use software for
prediction on other organisms is not available. A third class of methods makes direct use of protein structural data in order to predict whether the 3-dimensional conformations of proteins can dock with one another (Zhang et al. 2012), but this requires 3-dimensional protein structures that are often not available for non-model species.

The final class of PPI predictors attempts to combine the convenience of the interolog approach – namely the availability of sequence data – with the higher power of the various machine learning approaches. This may be accomplished by taking advantage of relatively short, conserved sequence motifs that either directly mediate interaction or are frequently associated with physical interaction. Several methods have use co-occurrence of protein domains to make predictions, including a random forest method that requires a single feature for each of the thousands of possible domains (Chen and Liu, 2005) and a genetic algorithm that uses existing PPI data to construct a set of domain-domain pairs that can be used for further PPI prediction (Singhal and Resat, 2007). While each of these may, in principle, allow for the prediction of physical interactions that cannot be readily detected through the interolog approach in the absence of functional data, the software for these methods has not been made available to the community. Moreover, neither method facilitates the incorporation of additional predictors such as subcellular localization or coexpression if such data is available.

**Insertion-Deletion (Indel) Polymorphism**

The gain or loss of sequence, termed an indel, is a common but understudied form of genetic variation in both prokaryotes and eukaryotes. Early comparisons of BAC sequences within species and between closely related species suggested that indel variation could account for majority of nucleotide changes (Britten et al. 2003), and more recent results confirm that indels are both abundant in populations and can form rapidly in both animals and plants (Mills et
al. 2006; DeBolt, 2010; Lu et al. 2012). Furthermore, the ability of indels to induce nonfunctionalization through frameshifts, deletions of entire exons, or the annihilation of key regulatory sequences may lead to insight into the process of fractionation in the paleopolyploid genome.

The mechanisms of indel formation are diverse and still under investigation. Small insertions are often generated through slippage of the DNA polymerase during replication (Nishizawa and Nishizawa, 2002), especially within microsatellite or other regions with simple repeat structures. Non-homologous end joining during double-stranded break repair may provide a second large source of small indels, particularly deletions, in both repetitive and non-repetitive regions (Montgomery et al. 2013). In contrast to small indels, large indels are primarily thought to arise through nonallelic recombination between homologous sequences with low-copy repetitive regions such as tandem gene arrays (Lupski and Stankiewicz, 2005; DeBolt, 2010), emphasizing their potential role in evolution through small-scale duplications. The mobilization of transposable elements may also result in the formation of large indels, and transposable elements are therefore found to be enriched in regions encompassed by both duplications and deletions (Bennetzen, 2000; DeBolt, 2010).

Although indels have received relatively little attention in plants, several studies suggest they may underlie a large percentage of functionally relevant variation. In humans, small indel diversity is diminished within both coding and non-coding functional regions relative to SNP variation, while it is also enriched within regions associated with expressional variation of known disease-causing genes (Montgomery et al. 2013). Likewise, transgressive and additive expression patterns in rice hybrids are associated with indel polymorphisms within promoter regions (Zhang et al. 2008). The transgressive patterns have particular relevance, as they may
provide a basis for the heterotic gains of hybrid individuals in agriculturally-important quantitative traits (Shivaprasad et al. 2012). Additional studies in *Arabidopsis thaliana*, *Glycine max*, and *Zea mays* suggest the importance of large indels toward local adaptation to pathogens and other forms of regional environmental stress, as evidenced by the enrichment of these indels in clusters of stress-responsive genes (Belo et al. 2010; Lu et al. 2012; McHale et al. 2012).

**Detecting indel variation**

Early genome-wide studies used microarray-based comparative genomic hybridization (CGH) to detect changes in copy number by comparing the hybridization signals of reference and population samples. While this is an effective method for discovering copy number variation in regions spanning at least 500 bp (Conrad et al. 2010; Park et al. 2010), CGH cannot reveal the genomic context of the duplications or deletions and is also relatively insensitive to single duplications compared to single deletions (Alkan et al. 2011). Fortunately, the introduction of cost-effective massively-parallel sequencing platforms now permits discovery and genotyping of localized small and large structural variants over entire genomes. This has led to the introduction of several algorithms designed to detect insertions, deletions, inversions, and transpositions using high-throughput short read data (Chen et al. 2009; Ye et al. 2009; Albers et al. 2011; Handsaker et al. 2011). Although the field has yet to reach consensus on best practices concerning indel detection and genotyping, most of these algorithms take advantage of the sequencing depth, the alignment of reads to the reference genome, the distance between paired end reads, or some combination thereof (Alkan et al. 2011).

Using a similar principle to CGH studies, methods using a read depth approach find regions where the sequencing depth significantly deviates from that expected under a diploid copy number. The read depth method also permits the inference of absolute copy number (Alkan
et al. 2009). However, its accuracy may be negatively impacted by biases that may be introduced during sequencing or assembly, which can cause high depth variance even in the presence of diploid copy numbers. Therefore most algorithms that exploit read depth have a method for modeling local depth variation (Klambauer et al. 2012).

In contrast to the read depth approach, the read pair and split read approaches each exploit discrepancies between individual query genome fragments and the reference genome. The split read approach accomplishes this by optimizing the alignment of short reads with the reference genome before finding gaps in either the reference or the read, which would be indicative of insertion or deletion, respectively (Alkan et al. 2011). The use of direct alignment allows the split-read method to identify the precise breakpoints of indels; however, it is limited by the size of the reads available, and the alignment of individual reads is highly computationally intensive (Albers et al. 2011). In the absence of large read lengths, the read pair approach can map large-scale structural variation by exploiting the difference between the expected distance between paired-end reads given the reference genome and the observed distance between paired-end reads given the insert size distribution. Under this scheme, paired ends that map to more distant locations in the reference genome than expected given the insert size are indicative of a deletion within the query sequence, while those that map to more proximal locations than expected suggest an insertion within the query sequence (Tuzun et al. 2005). However, this approach cannot determine the precise breakpoints for indels, and the size range of detectable indels largely depends upon the distribution of fragment sizes within the sequencing libraries (Medvedev et al. 2009).

The trade-offs between indel discovery methods have led several groups to develop software that combine two or more of the approaches. This strategy was taken by the 1000
Genomes Project to discover 50bp to 1Mb deletion polymorphisms across 168 genomes with an average of 4x coverage. Their approach, Genome STRiP, not only combines the read pair and read depth approaches during indel discovery but does so simultaneously within all individuals across a population (Handsaker et al. 2011). This population-based approach provides the algorithm with increased power to detect an indel, especially with low-coverage sequence. It also decreases the probability of false discovery due to sequencing artifacts that can mimic true structural variants (Handsaker et al. 2011).
References


Thomas, B. C., Pedersen, B., and Freeling, M., 2006. Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, **16**:934–946.


CHAPTER II

CONTRASTING PATTERNS OF EVOLUTION FOLLOWING WHOLE GENOME VERSUS TANDEM DUPLICATION EVENTS IN *POPULUS*
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 (Semon and Wolfe 2007; Lynch and Conery 2003). 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 (Jaillon et al. 2007; Ming et al. 2008; Schmutz et al. 2010; Schnable et al. 2009; The Arabidopsis Genome Initiative 2000; Tuskan et al. 2006). 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 (Chen and Ni 2006; Gaeta et al. 2007; 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 (Kohler et al. 2008; Leister 2004; Meyers et al. 2003). Unequal recombination is thought to be the primary mechanism driving the expansion of these gene clusters (Babushok et al. 2007; Leister 2004; 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 (Babushok et al. 2007; Zhang et al. 2005).
Following duplication, each gene within a paralagous pair may evolve in several ways. For example, it may retain the same set of functions as the ancestral copy (retention) (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 (Warren et al. 2010; Chapman et al. 2006; Hakes et al. 2007). 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 (Birchler and Veitia 2007, 2010; Edger and Pires 2009; Freeling and Thomas 2006). 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 (Davis and Petrov 2005; Maere et al. 2005; Seoighe and Gehring 2004). Meanwhile, the converse is true for other genes, such as those containing NBS-LRR motifs (Zhang et al. 2010; Leister 2004; Meyers et al. 2005, 2003).

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; (iii) determine the degree to which whole genome patterns of duplicate gene retention and expression conform to the expectations of the DDC hypothesis.
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 (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 and He 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 log2 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. 1-2). Although the microarray was based on gene models from Populus version 1.1 (Tsai et al. 2006), 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. 2008b). 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
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, $dN/dS$, the ratio of the nonnynonymous substitution rate to the synonymous substitution rate, was calculated for all duplicate pairs using the maximum likelihood method of Yoder and Yang (2007) implemented in the codeml module of PAML.

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 (Harrell 2009).

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:

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

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

where D denotes the data from the observed distribution of Spearman expression distances, \( \pi \)
values denote the mixture proportions of each distribution, and \( \mu \) and \( \sigma \) denote the means and
standard deviations of truncated normal distributions between 0 and 2. Initial parameter values
were drawn from uniform distributions, wherein means (\( \mu \)) were permitted to range between the
5th and 95th percentiles of the observed distribution, standard deviations (\( \sigma \)) were permitted to
range between 1/10 and the entire standard deviation of the observed distribution, and mixture
proportions (\( \pi \)) 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 (Dabney et al. 2010).
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 (Tang et al. 2008a; Tuskan et al. 2006). 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. We used logistic regression to identify significant predictors of duplicate retention following 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-r² = 0.07, P < 2e-16), and 12 were significant predictors of TD gene presence (Nagelkerke pseudo-r² = 0.121, P < 2e-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).
Size and Expression Breadth of gene duplicates

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 breadth than all other genes in the genome, while the opposite was true of TD genes (Table 2).

Functional Categories

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 TDs. However, closer inspection of the composition of these groups revealed substantial differences between tandem
and Salicoid duplicates (Supplementary Table 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 (Kohler et al. 2008; Meyers et al. 2003). 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 TDs (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 (Chen et al. 2006; Bassett et al. 2005). 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 TDs 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 TDs 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 TDs did not include any genes 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: $P = 2.074 \times 10^{-14}$), 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 polymorphisms and Expression Imbalance

Salicoid duplicates had a significantly lower ratio of nonsynonymous to synonymous substitutions ($dN/dS$) in coding regions compared to TD genes (Table 2). This was true over a range of $dS$ values (Supplemental Fig. 6-7). Furthermore, using a forward selection strategy, we found a nonlinear relationship between $dN/dS$ 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 $dN/dS$ 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 $dN/dS$ and expression distance was most evident at high combined breadth and low expression imbalance (Fig. 5).
Table 2.1 Significant predictors for retention of genes in Salicoid and tandem duplicate pairs

<table>
<thead>
<tr>
<th></th>
<th>Salicoid Duplicates</th>
<th></th>
<th></th>
<th>TDs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>exp(β)</td>
<td>SE</td>
<td>z</td>
<td>P(&gt;</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>-0.57</td>
<td>0.57</td>
<td>0.03</td>
<td>-20.71</td>
<td>0.00</td>
</tr>
<tr>
<td>Tandem</td>
<td>-1.35</td>
<td>0.26</td>
<td>0.08</td>
<td>-17.26</td>
<td>0.00</td>
</tr>
<tr>
<td>Salicoid</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Breadth</td>
<td>0.57</td>
<td>1.77</td>
<td>0.04</td>
<td>15.90</td>
<td>0.00</td>
</tr>
<tr>
<td>Gene Length</td>
<td>0.10</td>
<td>1.11</td>
<td>0.01</td>
<td>16.90</td>
<td>0.00</td>
</tr>
<tr>
<td>Protein Binding</td>
<td>0.29</td>
<td>1.34</td>
<td>0.04</td>
<td>6.58</td>
<td>0.00</td>
</tr>
<tr>
<td>Transmembrane</td>
<td>-0.10</td>
<td>0.90</td>
<td>0.03</td>
<td>-3.83</td>
<td>0.00</td>
</tr>
<tr>
<td>Transcription Factor</td>
<td>0.66</td>
<td>1.93</td>
<td>0.09</td>
<td>7.18</td>
<td>0.00</td>
</tr>
<tr>
<td>Catalytic</td>
<td>-0.19</td>
<td>0.83</td>
<td>0.03</td>
<td>-6.10</td>
<td>0.00</td>
</tr>
<tr>
<td>Receptor</td>
<td>-1.49</td>
<td>0.23</td>
<td>0.28</td>
<td>-5.31</td>
<td>0.00</td>
</tr>
<tr>
<td>Stress</td>
<td>-0.74</td>
<td>0.48</td>
<td>0.09</td>
<td>-8.13</td>
<td>0.00</td>
</tr>
<tr>
<td>Kinase</td>
<td>0.25</td>
<td>1.28</td>
<td>0.06</td>
<td>4.35</td>
<td>0.00</td>
</tr>
<tr>
<td>Phosphatase</td>
<td>0.70</td>
<td>2.01</td>
<td>0.15</td>
<td>4.61</td>
<td>0.00</td>
</tr>
<tr>
<td>Nucleic Acid Binding</td>
<td>0.15</td>
<td>1.16</td>
<td>0.04</td>
<td>3.34</td>
<td>0.00</td>
</tr>
<tr>
<td>Transporter</td>
<td>0.17</td>
<td>1.19</td>
<td>0.06</td>
<td>2.70</td>
<td>0.00</td>
</tr>
<tr>
<td>Ligase</td>
<td>0.61</td>
<td>1.84</td>
<td>0.11</td>
<td>5.36</td>
<td>0.00</td>
</tr>
<tr>
<td>Protease</td>
<td>0.24</td>
<td>1.27</td>
<td>0.09</td>
<td>2.65</td>
<td>0.01</td>
</tr>
<tr>
<td>Cation Binding</td>
<td>0.13</td>
<td>1.14</td>
<td>0.04</td>
<td>3.33</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 2.2 Mean, median, and standard error for the gene length, expression breadth, and $d_N/d_S$ (of pairs) within the whole genome, among Salicoid duplicates, and among TDs. P-values for gene length and breadth are based on van der Waerden tests for independence of gene characteristic and duplicate type.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Median</th>
<th>SE</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole Genome</td>
<td>2.587 kb</td>
<td>2.007 kb</td>
<td>0.0126</td>
<td></td>
</tr>
<tr>
<td>Salicoid Duplicates</td>
<td>2.854 kb</td>
<td>2.304 kb</td>
<td>0.0183</td>
<td>0.0000</td>
</tr>
<tr>
<td>TDs</td>
<td>2.126 kb</td>
<td>1.767 kb</td>
<td>0.0471</td>
<td>0.0000</td>
</tr>
<tr>
<td>Breadth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole Genome</td>
<td>0.5396</td>
<td>0.5714</td>
<td>0.00185</td>
<td></td>
</tr>
<tr>
<td>Salicoid Duplicates</td>
<td>0.5774</td>
<td>0.6428</td>
<td>0.00264</td>
<td>0.0000</td>
</tr>
<tr>
<td>TDs</td>
<td>0.3868</td>
<td>0.2857</td>
<td>0.00875</td>
<td>0.0000</td>
</tr>
<tr>
<td>$d_N/d_S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salicoid Duplicates</td>
<td>0.2728</td>
<td>0.2428</td>
<td>2.29e-5</td>
<td></td>
</tr>
<tr>
<td>TDs</td>
<td>0.4478</td>
<td>0.40295</td>
<td>6.57e-4</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3 Proportion of genes annotated with a given functional category in each of the 2 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Mix. Dist. 1</th>
<th>Mix. Dist. 2</th>
<th>p-value</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Binding</td>
<td>0.11</td>
<td>0.11</td>
<td>0.92</td>
<td>0.51</td>
</tr>
<tr>
<td>Transmembrane</td>
<td>0.28</td>
<td>0.31</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>0.04</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Catalytic</td>
<td>0.29</td>
<td>0.32</td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>Receptor</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Ion</td>
<td>0.00</td>
<td>0.00</td>
<td>0.53</td>
<td>0.38</td>
</tr>
<tr>
<td>Stress</td>
<td>0.01</td>
<td>0.01</td>
<td>0.11</td>
<td>0.14</td>
</tr>
<tr>
<td>Kinase</td>
<td>0.05</td>
<td>0.05</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td>Phosphatase</td>
<td>0.01</td>
<td>0.01</td>
<td>0.55</td>
<td>0.38</td>
</tr>
<tr>
<td>Nucleic Acid Binding</td>
<td>0.14</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Transporter</td>
<td>0.04</td>
<td>0.04</td>
<td>0.71</td>
<td>0.44</td>
</tr>
<tr>
<td>Ligase</td>
<td>0.02</td>
<td>0.02</td>
<td>0.75</td>
<td>0.44</td>
</tr>
<tr>
<td>Protease</td>
<td>0.03</td>
<td>0.02</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Cation binding</td>
<td>0.12</td>
<td>0.13</td>
<td>0.47</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Table 2.4 Regression table for the relationship between expression distance and $d_{S}/d_{S}$, 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   |
| $d_{S}/d_{S}$ | 0.2607   | 0.1829     | 1.43    | 0.1540   |
| $(d_{S}/d_{S})^2$ | -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   |
| $d_{S}/d_{S}$:Combined Breadth | 0.5982   | 0.1907     | 3.14    | 0.0017   |
| $d_{S}/d_{S}$:Exp. Imbalance | -0.1903 | 0.0281     | -6.78   | 0.0000   |
| $d_{S}/d_{S}$:Combined Breadth:Exp. Imbalance | 0.1324 | 0.0313     | 4.23    | 0.0000   |
Figure 2.1 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 2.2 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 2.3 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 2.4 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 2.5 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.
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. This pattern is consistent with the stochastic degeneration of subfunctions predicted under the DDC model. 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 a dominant 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 (Woodhouse et al. 2010; Kong et al. 2007; Zhang 2003; Freeling 2009). Each of these may result in the incomplete duplication of genes in the array, 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 dN/dS observed between TDs is consistent with weaker purifying selection for the retention of ancestral gene functions.
Furthermore, the proportion of TDs for which there was evidence of positive selection (i.e. $dN/dS > 1$) was an order of magnitude higher than for Salicoid duplicates ($P = 7.75 \times 10^{-10}$). 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 TDs 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 (Afzal et al. 2008; Shiu et al. 2004). 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 (Lehti-Shiu et al. 2009; Mondragon-Palomino et al. 2002).

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

Lastly, we observed that TDs tend to display much more extreme patterns of functional overrepresentation than Salicoids. This illustrates an important distinction between the modes of duplication. While the set of Salicoid duplicates reflects the pattern of retention following a
single duplication event, the tandemly duplicated genes reflect both the propensity of certain genes to duplicate in tandem and the rate of elimination by selection or drift. In the case of R-genes, stress is known to increase rates of somatic recombination, which can facilitate the birth of new genes through unequal crossing over if it occurs within reproductive cell lineages (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 (Birchler and Veitia 2010; Aury et al. 2006). 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 dN/dS 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 dN/dS. 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 substitutions 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 (Ettwiller and Veitia 2007; Pal et al. 2001). This is consistent with the negative correlation we observed between combined breadth and dN/dS. 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 dN/dS 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 consistent with random partitioning of subfunctions, suggesting many whole genome duplicates are not subject to constraints to maintain 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.
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, Grants No. DE-FG02-06ER64185 and DE-FG02-07ER64449 and by funding from the BioEnergy Science Center, a U.S. DOE Bioenergy Research Center (Office of Biological and Environmental Research in the DOE Office of Science).

Statement of Original Publication

This work was originally published in the January, 2012 edition of Genome Research (Rodgers-Melnick et al. 2012).
References


Kohler, A., Rinaldi, C., Duplessis, S., Baucher, M., Geelen, D., Duchaussoy, F., Meyers, B.C.,
Boerjan, W., and Martin, F. 2008. Genome-wide identification of NBS resistance genes in

Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for


Leister, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of

Genetics* **38**: 124-129.


Biology* **7**: 213.


Supplemental Figure 2.1 Boxplots for the $d_s$ ranges in Salicoid and tandem duplicate pairs
Supplemental Figure 2.2 Mean $d_N/d_S$ for Salicoid and tandem duplicates grouped into bins of 0.2 $d_S$. Error bars indicate standard error of the mean.
Supplemental Figure 2.3 Scatter plots of the replicates for the slides used in this analysis. Each plot contains a line of best fit based on least squares regression, along with the R-squared value.
Supplemental Figure 2.4 Plots of the replicates for two of the slides used in this analysis that had relatively poor correspondence between replicates. These were used for further analysis because they clustered with other floral tissues (Supplemental Figure 5).
Supplemental Figure 2.5 Hierarchical clustering based on the Euclidean distance for all replicates in the original microarray. VYL, SE, ApL, and Xpi slides were dropped for subsequent analyses.
CHAPTER III

PREDICTING WHOLE GENOME PROTEIN INTERACTION NETWORKS FROM PRIMARY SEQUENCE DATA IN MODEL AND NON-MODEL ORGANISMS USING ENTS
Abstract

**Background:** The large-scale identification of physical protein-protein interactions (PPIs) is an important step toward understanding how biological networks evolve and generate emergent phenotypes. However, experimental identification of PPIs is a laborious and error-prone process, and current methods of PPI prediction tend to be highly conservative or require large amounts of functional data that may not be available for newly-sequenced organisms.

**Results:** In this study we demonstrate a random-forest based technique, ENTS, for the computational prediction of protein-protein interactions based only on primary sequence data. Our approach is able to efficiently predict interactions on a whole-genome scale for any eukaryotic organism, using pairwise combinations of conserved domains and predicted subcellular localization of proteins as input features. We present the first predicted interactome for the forest tree *Populus trichocarpa* in addition to the predicted interactomes for *Saccharomyces cerevisiae*, *Homo sapiens*, *Mus musculus*, and *Arabidopsis thaliana*. Comparing our approach to other PPI predictors, we find that ENTS performs comparably to or better than a number of existing approaches, including several that utilize a variety of functional information for their predictions. We also find that the predicted interactions are biologically meaningful, as indicated by similarity in functional annotations and enrichment of co-expressed genes in public microarray datasets. Furthermore, we demonstrate some of the biological insights that can be gained from these predicted interaction networks. We show that the predicted interactions yield informative groupings of *P. trichocarpa* metabolic pathways, literature-supported associations among human disease states, and theory-supported insight into the evolutionary dynamics of duplicated genes in paleopolyploid plants.

**Conclusion:** We conclude that the ENTS classifier will be a valuable tool for the *de novo* annotation of genome sequences, providing initial clues about regulatory and metabolic network topology, and revealing relationships that are not immediately obvious from traditional homology-based annotations.
**Introduction**

Proteins do not exist within a vacuum. Much of the startling diversity of living organisms emerges only with the aggregate combinatorial complexity of protein-protein interactions (PPIs) (Lander et al., 2001). As such, the discovery of physical interactions between proteins is often an essential step in the characterization of protein functions, providing insights into diverse cellular processes such as the fluxes of metabolic pathways, the logic of transcriptional activation, and the kinetics of signal transduction. This in turn forms the basis of understanding biological functions at the organismal scale, including mechanisms of environmental responses and the etiology of disease states.

Despite the biological importance of PPIs and the availability of high-throughput screening methods in recent years, experimentally-verified PPI networks remain sparsely populated, especially with respect to the amount of sequence data currently available. High throughput approaches such as automated yeast two-hybrid screens and tandem affinity purification/mass spectrometry have detected thousands of binary PPIs in animal and fungal model organisms such as *Homo sapiens* (Stelzl et al., 2005), *Saccharomyces cerevisiae* (Yu et al., 2008), and *Drosophila melanogaster* (Guruharsha et al., 2011), yet the current size of the interactome belonging to the experimental workhorse of the plant kingdom, *Arabidopsis thaliana*, only constitutes approximately 3\% of its expected size (Consortium, 2011). Moreover, the lack of significant numbers of PPIs for non-model species hinders the development of evolutionary studies concerning rewiring within the interactome (De Smet and Van de Peer, 2012).

The demand for additional PPIs has led to the development of several methods for computational PPI prediction over the past decade. Several groups have attempted to expand the *A. thaliana* interactome using statistical learning methodology and/or transfer of interaction
annotation based on homology (interologs) (De Bodt et al., 2009; Cui et al., 2008; Geisler-Lee et al., 2007; Lin et al., 2011). Similar methods have been used to expand the number of network connections in the S. cerevisiae and H. sapiens proteomes. Though each of these methods does have the potential to provide useful network information, each approach carries distinct disadvantages limiting its use on non-model species. The interolog-based approaches are limited to discovering PPIs for only the most conserved proteins, as reflected by their relatively low discovery rates in A. thaliana (Geisler-Lee et al., 2007). Other methods rely on an ensemble of functional data, such as genome-wide measures of co-expression and co-localization, which is often not available for non-model organisms.

Although many homologous proteins may evolve to become highly dissimilar at the primary sequence level, they often retain conserved structural and/or functional units known as domains. These domains may directly mediate interactions between proteins, as demonstrated by databases of domain-domain interactions such as DOMINE (Yellaboina et al., 2011). However, even in the absence of direct interaction, certain pairwise combinations of domains suggest a high probability of interaction. Thus, domain-based approaches of PPI prediction have the potential to provide the advantages of the interolog-based approaches while maintaining utility for less-conserved proteins, especially if the approach also includes features more sensitive to fine-scale differences in amino acid content. There have been several recent attempts to infer PPIs based on pairwise domain information. For example, Singhal et al. used a genetic algorithm to discover domain-domain interactions that could be used as predictors of PPIs (Singhal and Resat, 2007). Chen and Liu used a domain-driven random forest classifier to predict PPIs for S. cerevisiae (Chen and Liu, 2005). However, feature representation for the algorithm required vectors with several thousand entries, making its use computationally expensive for full-genome
prediction. Although these approaches are initially promising, there is a lack of publicly-available software that would enable domain-based PPI prediction on a genome-wide scale for non-model organisms lacking large experimental data sets.

Here we present "Elucidating Network Topology with Sequence" (ENTS), a binary PPI classifier that uses a random forest framework. ENTS is capable of efficiently and exhaustively evaluating all potential protein-protein pairs in a large eukaryotic genome using parallelization. We show that the method provides comparable or better predictions on recently experimentally-determined PPIs than several existing methods and that such predictions are biologically plausible using the predicted interactomes of A. thaliana, P. trichocarpa, M. musculus, H. sapiens, and S. cerevisiae. Scripts, instructions for use, and predicted PPIs for several organisms are available at http://ENTS.as.wvu.edu.
Methods

Data sources

We obtained experimentally verified physical interactions for *H. sapiens*, *S. cerevisiae*, *M. musculus*, *D. melanogaster*, and *A. thaliana* from the PPI databases DIP (10/27/2011 Release) (Xenarios et al., 2002), IntAct (01/01/2012 Release) (Kerrien et al., 2012), and BioGRID (v. 3.1.84) (Stark et al., 2011). Additional experimentally-verified interactions for *A. thaliana* were taken from the TAIR database (05/27/2009 Release) (Lamesch et al., 2012). We also collected known and predicted domain-domain interactions and their associated confidence scores from the DOMINE database (version 2) (Yellaboina et al., 2011). All protein annotations were taken from Ensembl core databases with the *A. thaliana* and *P. trichocarpa* annotations corresponding to Ensembl Plants release 12, the *S. cerevisiae* annotations corresponding to Ensembl Fungi release 12, and the *H. sapiens*, *M. musculus*, and *D. melanogaster* annotations corresponding to Ensembl release 65. Only canonical versions of proteins were used; splice variants were not considered in this analysis.

Calculation of domain pair odds

We calculated a log-of-odds score for each pair of domains observed at least once in an interacting protein pair. This score may be interpreted as the odds of observing a pair of domains in an interacting protein pair versus by random chance among all interacting proteins. We obtained all unique PFAM domains present in each protein for all experimentally-verified protein interactions in *H. sapiens*, *S. cerevisiae*, *M. musculus*, *D. melanogaster*, and *A. thaliana*. We then assessed all possible pairwise domain combinations among all these protein pairs and calculated the log-odds score for each domain pair as follows:

\[
f(D_x, D_y) = \frac{n(D_x, D_y)}{\sum_{i=1}^{n_p} \sum_{j=1}^{l} n(D_i, D_j)}
\]
where \( n(D_x,D_y) \) is the number of times the domain pair \( D_x, D_y \) was observed among experimentally-verified protein interactions, \( n_p \) is the total number of domain pairs, and \( n(D_x) \) is the number of times \( D_x \) was observed among the set of proteins with experimentally-verified interactions. Because the absence of a protein pair within the experimentally-verified data set may result from either non-detection of existing interactions or the absence of any such interactions, we chose to assume an LOD of 0 for all domain pairs that were not observed in any interacting protein pairs.

**Feature data**

All data features were defined from pairwise-domain information and predictions of subcellular localization. An exhaustive list of all features can be found in supplementary table 1. Domain-based features included the sum of all odds scores, the highest odds score, the lowest odds score (ceiled at 0), the number of pairwise domain pairs not observed in any of the interacting proteins used to calculate the odds scores, the number of domain pairs found among the pairs with odds scores, the number of domain pair predicted or known to interact in DOMINE, and the highest DOMINE confidence score assigned to a domain pair. All other features were outputs of the high-res MultiLoc2 subcellular predicted program (Blum et al., 2009), which used protein sequence data and computer-generated GO categories as input. These features included the probabilities of localization to each of the possible subcompartments - cytoplasm, nucleus, peroxisome, ER, mitochondria, chloroplast (plant only), vacuole (plant and
fungus), and the lysosome (animal only) - along with the raw output from each of the MultiLoc2 subprograms - SVMTarget, SVMSA, SVMaac, PhyloLoc, GOLoc, and MotifSearch.

*Training and prediction*

Classifiers were trained using 1,330 randomly selected protein pairs with experimental verification of interaction to serve as positive examples for each organism. We also included 101,300 randomly drawn pairs of proteins without any known or predicted interactions in the Reactome version 39 database to serve as negative examples. In order to avoid potential overfitting based on the predicted subcellular characteristics of proteins in the positive set, we spiked the negative set with 1,300 randomly selected pairs of proteins without known or predicted interactions drawn from the positive set, wherein the proteins in the pair were drawn in proportion to their representation within the interacting pairs. Additionally, we included each protein pair twice within the dataset but switched the subcellular features between the proteins in the second set (reversed set). We trained the classifier using 400 fully-grown trees with the R randomForest package (http://cran.r-project.org/package=randomForest). Prediction also used both the forward and reversed set, with the final predictions taken from the union of the two. We trained a total of 3 ENTS classifiers - one on *H. sapiens*, one on *S. cerevisiae*, and one on *A. thaliana*. The *H. sapiens*-trained classifier was used to predict on *H. sapiens* and *M. musculus*; the *S. cerevisiae*-trained classifier was used to predict on *S. cerevisiae*; and the *A. thaliana*-trained classifier was used to predict on *A. thaliana* and *P. trichocarpa*.

In order to characterize the testing performance, we calculated the area under the ROC curve (AUC) using the R pROC package (http://cran.r-project.org/web/packages/pROC/index.html). Testing data consisted of 1,300 known positive interactions and 101,300 randomly selected negative examples without known or predicted
interactions. Testing data did not include any protein pairs used to calculate the domain pair odds and contained no overlap to the random forest training data at the level of protein interactions.

Comparisons to other predictors

Predicted protein interaction datasets were taken from the Pitre et al. PIPE2 *S. cerevisiae* classifier novel interactions (Pitre et al., 2006), the *H. sapiens* PIPS classifier (McDowall et al., 2009), the NIA Mouse Protein-Protein Interaction Database (Yellaboina et al., 2008), the predicted *A. thaliana* interactome of De Bodt et al. (De Bodt et al., 2009), the AtPID *A. thaliana* interactome (Cui et al., 2008), the Geisler-Lee *A. thaliana* interactome (Geisler-Lee et al., 2007), and the PAIR high-confidence *A. thaliana* interactome (Lin et al., 2011b). We obtained protein interaction datasets without any overlap to the interactions used to train ENTS or calculate domain odds scores. We also limited the overlap to training data from the alternative classifiers by choosing datasets containing PPIs that were either published after the alternative classifier's publication date or were not present within the datasets reportedly used for training. In *S. cerevisiae* these were taken from a high-throughput yeast two-hybrid study (Yu et al., 2008). In humans we used high confidence sets of human protein interactions from a recent study of protein complexes (confidence of at least 0.9) (Havugimana et al., 2012) and a high-throughput mass-spectrometry study of human PPIs (confidence of at least 0.3) (Ewing et al., 2007). In *M. musculus* we used interactions taken from the HitPredict database (05/01/2012) that were published after October 2008 (Patil et al., 2011). Finally, in *A. thaliana* we used high-throughput yeast two-hybrid and literature curated interactions from a large study of network evolution (Consortium, 2011).

Prior to performing each comparison we removed any predicted pairs from ENTS and the alternative classifiers that were used for training of the ENTS random forest or calculation of
domain pair odds. ENTS predictions were then sorted by decreasing confidence and truncated to sets of the same size as the alternative classifiers. We then calculated the sizes of the intersections with the experimental datasets in addition to the sizes of the three-way intersections between ENTS, the alternative classifiers, and the experimental datasets.

**Evaluation of functional similarity**

For each measure of functional similarity, we first narrowed the protein network (predicted or experimentally-verified) to the set of proteins with the given annotation present. For instance, we narrowed the networks to those with at least 1 KEGG annotation when evaluating KEGG similarity. Similarity was then calculated as the number of pathways present in the intersection of two proteins’ annotations divided by the number of pathways in the union.

GO semantic similarity between pairs of putatively interacting proteins was calculated separately for biological process, cellular component, and molecular function, and was based on the information content of shared parents (Lord et al., 2003). Briefly, for each organism we assigned a probability to each GO subgraph node, $p(c)$, which was defined as follows:

$$p(c) = \frac{n(c_d)}{n_c}$$

where $n(c_d)$ is the number of times the node or any of its descendants occurred in the genome, and $n_c$ is the number of times any term occurred. For each pair of queried terms we then found the shared parent with the minimal probability $p_m(c_x, c_y)$ among the set of shared parents in the subgraph, $S(c_x, c_y)$:

$$p_m(c_x, c_y) = \min_{c \in S(c_x, c_y)} p(c)$$
The similarity score between the two terms $c_x$ and $c_y$ was then defined as follows:

$$\text{sim}(c_x, c_y) = -\ln\left(p_m(c_x, c_y)\right)$$

In the case of more than one pair of GO terms between a pair of proteins, we set the similarity score as the maximum of all pairwise similarity scores.

*Metabolic pathway linkages*

We made a network of linkages between metabolic pathways downloaded from the PoplarCyc v. 3.0 database (Zhang et al., 2010) using protein interactions predicted in the *P. trichocarpa* 0.65 confidence network. Pathways were linked to one another if one or more protein pairs between the two pathways were predicted to interact. We also did not include predicted self-interactions when constructing the network. We tested the significance of all pathway linkages by creating 10,000 randomized PPI networks in which each node had the same degree as the ENTS network but randomized connections. We retained pathway linkages for which the number of supporting interactions was greater than that of at least 99.9% of randomized networks. We found the number of pathway linkages that connected two pathways sharing at least one metabolic compound between them, discounting all compounds that were present in more than 15 pathways in order to prevent less meaningful associations due to common compounds such as ATP. We then assessed the significance of the compound sharing by generating 10,000 randomized pathway linkage networks in which each pathway had the same degree as in the observed network.

We performed clustering on the pathway linkage network using the MCL graph clustering algorithm with default parameters (van Dongen, 2000). Edges were weighted by the fraction of
interactions predicted to exist between the two pathways out of the total number of possible non-
self interactions. We then tested the clusters for significant enrichment using annotation
enrichment analysis, which corrects for biases that can occur under Fisher's exact test (Glass and
Girvan, 2012). Briefly, each pathway was contained within a pathway ontology retrieved from
PlantCyc, forming a directed acyclic graph. Each node in the graph was annotated with a given
pathway if that pathway was a descendant of the node. P-values were then generated for each
term in the ontology as follows:

\[ p(M_{gt}) = \sum_{i=M_{gt}}^{\min(M_g,M_t)} \frac{(\frac{M_i}{M_{gt}})(M_{tot}-M_{gt}-i)}{M_{tot}} \]

where \( M_g \) is the number of ontology annotations to the cluster, \( M_{gt} \) is the number of ontology
annotations to the cluster on the ontology branch of interest, \( M_t \) is the number of pathways
annotated to the branch of interest, and \( M_{tot} \) is the total number of pathway annotations made to
the ontology graph. We assessed significance using a 0.05 family-wide type I error rate under a
Bonferroni correction, such that a term was considered significant if \( p \leq \frac{0.05}{n_c n_o} \), where \( n_c \) is the
number of clusters found and \( n_o \) is the number of nodes in the ontology.

**OMIM disease network**

We created a network of OMIM diseases and disorders by creating an edge between two
diseases if their underlying causative loci in the OMIM database were predicted to interact
within the ENTS predicted human network. We did not count self-interactions when creating the
network. We tested the significance of the disease-disease associations by creating 10,000
randomized PPI networks in which each node had the same degree as the ENTS network but
randomized connections. We then retained disease associations for which the number of supporting interactions was greater than that of at least 99.9% of randomized networks.

We then evaluated the significance of the network associations using Pubmed literature mining. In order to perform this with a controlled vocabulary, we first mapped the OMIM identifiers for each disease to medical subject headings (MeSH) terms using the Gendoo database (Nakazato et al., 2009), narrowing the disease network to OMIM identifiers associated with at least 1 MeSH term. Each of these OMIM-MeSH associations was associated with a p-value indicating the significance of the term-term association, \( p_{OM} \). We obtained the full set of Pubmed IDs (PMIDs) associated with each MeSH term present within the disease network using NCBI E-Utilities. We then assessed literature similarity using a measure that incorporated both the significance of the OMIM-MeSH association and the frequency of the MeSH-MeSH pairings within the literature. This was calculated as follows:

\[
\text{sim}(M_x, M_y) = -\frac{n(P_x \cap P_y)}{\min(n(P_x), n(P_y))} \log \left( \max \left( p_{OM}(M_x), p_{OM}(M_y) \right) \right)
\]

where the \( n(.) \) function refers to the number of PMIDs in the set, and \( P_x, P_y \) are the sets of PMID ids associated with the MeSH terms \( M_x \) and \( M_y \). Only the most significant MeSH term for each OMIM disease was used for comparison. In order to assess the significance of the observed network, we compared the distribution of literature scores to those in 250 random disease-disease networks in which disease nodes were sampled in proportion to their degree in the observed network, thus generating networks with similar degree distributions.

*WGD logistic regression*
Intragenomic syntenic segments corresponding to the Salicoid WGD in *P. trichocarpa* and the α and β/γ duplications in *A. thaliana* were defined using MCScan (Tang et al., 2008). We separated the more recent α-duplication from the older β and γ duplications using the mappings provided by Bowers et al (2003). However, many of the duplicated regions with higher dS were not defined within that dataset, and we could not confidently separate the β and γ duplications from one another based on dS due to saturation of substitutions. Therefore, we considered the *A. thaliana* β and γ WGDs as a single β/γ duplication for the purposes of the analysis. We then fit the following model using the generalized linear model with the logit link function in the R programming language:

\[
\text{Duplication} \sim \log(\text{Degree Centrality}) + \text{Duplicated Neighbors} + \log(\text{Degree Centrality}) \times \text{Duplicated Neighbors}
\]

The response was coded as 0 or 1, corresponding to absence or presence of a duplicate paralog from the given WGD, respectively. The duplicated neighbors term was calculated as the fraction of neighboring genes that were retained as duplicates from the corresponding WGD.

Degree centrality was calculated as follows using the Python Networkx package (http://networkx.lanl.gov):

\[
\text{Degree Centrality} = \frac{d(n_x)}{n_G - 1}
\]

where \(n_G\) is the number of nodes in the graph, and \(n_x\) is the degree of node \(n_x\). We assessed the significance of the model coefficients by resampling 10,000 times with replacement, fitting the model to the resampled data, and generating 95%, 99%, and 99.9% confidence intervals using the corresponding quantiles of the coefficient distributions. Additionally, because the degree centrality constrained the possible range of the duplicated neighbor fraction such that a gene with one neighbor could only have a duplicated neighbor value of either 0 or 1, we fit two models for
each duplication: one using the full set of genes with at least 1 interaction in the predicted network and one using only the subset with at least 10 predicted interactions.
Results

ENTS performance relative to experimental predictions

We assessed the performance of ENTS by calculating the area underneath the ROC curve (AUC) for testing data consisting of no overlap with the training data at the level of protein interaction and no overlap with any protein pairs used to calculate pairwise domain LOD scores (see methods). AUC scores ranged from 0.811 and 0.827 in the yeast and human-trained classifiers, respectively, to a high of 0.9632 in the A. thaliana-trained classifier (Fig. 1). The most important features for the classifier included those derived from the analysis of domain pairs and the output from the subcellular localization prediction program multiLoc2 that measures the extent to which the amino acid contents and phylogenetic profiles of the query proteins matched particular cellular compartments (Additional file 1; Supplemental Figs. 5-7).

ENTS performance relative to other classifiers

We obtained whole-genome predictions of PPIs for the organisms on which the classifiers were trained (i.e., S. cerevisiae, H. sapiens and A. thaliana), as well as for species that were not used in training the predictors. We find that the numbers of predicted interactions are highly similar between the training and prediction species, although the training species do have an enrichment of genes at the high confidence levels (Table 1).

In order to assess the performance of ENTS on whole genome data relative to several alternative classifiers, we examined the frequency of positive predictions among sets of experimentally-determined PPIs that were not used for training, testing, or calculation of pairwise domain odds in ENTS or used for training in the alternative classifiers. We obtained genome-wide predicted PPI datasets in S. cerevisiae (Pitre et al., 2006), H. sapiens (McDowall
et al., 2009), *M. musculus* (Yellaboina et al., 2008), and *A. thaliana* (De Bodt et al., 2009; Cui et al., 2008; Geisler-Lee et al., 2007; Lin et al., 2011b). The sizes of the predicted datasets varied greatly, so when making comparisons we reduced the sizes of the ENTS predictions to those of the alternative datasets following removal of predictions that corresponded to data used for ENTS training or calculation of pairwise domain odds.

For *A. thaliana*, we compared predictions to 6,314 novel yeast two-hybrid and literature-curated PPIs from a large-scale study of interactome evolution (Consortium, 2011). We found that ENTS predicted more of these interactions than 3 of the 4 alternative classifiers (Fig. 2a). This included more than twice as many predicted interactions as the Geisler-Lee (n=19,779) and De Bodt (n=51,594) sets, each of which used interolog approaches to make their predictions (De Bodt et al., 2009; Geisler-Lee et al., 2007). By contrast, ENTS made a similar number of positive predictions to the AtPID classifier (n=24,248) and less positive predictions than the PAIR classifier (n=143,939). Those two approaches used machine learning techniques - naive Bayes and SVM, respectively - to combine interolog data with domain content and functional data such as co-expression, gene ontology similarity, and co-localization (Cui et al., 2008; Lin et al., 2011a). Strikingly, each classifier shared relatively few of its positive predictions with the ENTS predictor, with the highest number of interactions shared between the ENTS and PAIR classifiers, at 36.8% of the ENTS predictions.

We find that the number of novel interactions predicted by ENTS is similar to those predicted by the Pitre et al. PIPE2 classifier (n=13,826) in *S. cerevisiae* and the PIPS classifier (n=22,687) in human (Pitre et al., 2006; McDowall et al., 2009) (Fig. 2b-c), as found through comparisons to a high-throughput yeast two-hybrid dataset in *S. cerevisiae* (Yu et al., 2008) (n=1,337) and high-confidence interactions from two large-scale studies of human PPIs
(n=2,045) (Ewing et al., 2007; Havugimana et al., 2012). Again, relatively few predictions were shared between each pair of classifiers. The alternative classifiers differed substantially in their prediction methods. The Pitre et al. classifier based its predictions on the pairwise-occurrence of short sequence motifs (Pitre et al., 2006), while the PIPS classifier used naive Bayes to combine sequence-derived features such as orthology and pairwise domain content with functional data such as gene co-expression, post-translational modifications, and co-localization (McDowall et al., 2009). In mouse, we find that ENTS predicted many fewer novel interactions from a high-confidence set of literature-curated interactions (n=1,807) relative to the interolog-based predictor of Yellaboina et al. (n=36,608) (Fig. 2d), despite its having 2.3 times the number novel interactions predicted in humans within a dataset 1.6 times the size of the PIPS set (Yellaboina et al., 2008). The high number of mouse interologs inferred directly from human interactions is responsible for most of this disparity, as ENTS predicts more of the novel interactions (36 vs. 25) when the human-derived interologs are filtered out of the Yellaboina et al. dataset.

**Biological plausibility of PPI predictions**

The majority of predicted interactions for each organism are not experimentally verified, so we required indirect means of assessing their plausibility. This led us to assess the similarity of annotations and expression profiles among predicted interactors, excluding self-interactions to avoid upward bias. We find that ENTS-predicted interactors share KEGG pathways significantly more often than expected by chance for all species (Fig. 3a). Moreover, KEGG similarity for the organisms with the largest experimentally-determined interactomes - H. sapiens and S. cerevisiae - matches or exceeds those of the experimentally-verified networks (Fig. 3c). M. musculus KEGG similarity closely matches that observed for H. sapiens, while the measures for P. trichocarpa actually exceed those of A. thaliana at higher confidence levels (Fig. 3a). We also
find that the mean semantic similarities between predicted interactors for GO biological process (BP), GO cellular component (CC), and GO molecular function (MF) are significantly greater than expected by chance (Fig. 3b, Additional File 1; Supplemental Fig. 4). However, even though GO categories were not included as predictors during random forest prediction, they were used during subcellular localization prediction, so their use as a verification criterion is somewhat circular.

Lastly, we find that the distributions of Pearson gene expression correlations between ENTS-predicted interactors are significantly enriched for co-expressed genes ($\rho > 0.5$) in all organisms (Fig. 4). The extent of enrichment varies by organism, with all confidence levels yielding significant enrichment of co-expressed genes in *S. cerevisiae*, *H. sapiens*, and *A. thaliana*. ENTS produces a significant enrichment of co-expressed genes for *M. musculus* at 0.55 and 0.65 confidence levels and for *P. trichocarpa* at all confidence levels except 0.85. Notably, however, the *M. musculus* experimentally-verified network is not significantly enriched for co-expressed genes under the microarray experiment used. Several alternative prediction methods yield networks with much higher co-expression than those predicted by ENTS. However, several of these - PAIR, AtPID, and PIPS - used co-expression as a predictor of protein interaction (Lin et al., 2011b; Cui et al., 2008; McDowall et al., 2009).

Metabolic pathway linkages in *P. trichocarpa*

The connectivity of PPI networks permits insight into higher-order structures that largely remain hidden under non network-based analyses. Several studies have demonstrated that biological networks are organized into pathways or modules, each of which contain highly connected groups of genes that may act semi-autonomously with respect to the action of the
network as a whole (Wagner et al., 2007). We used the predicted *P. trichocarpa* PPI network to infer a network of *P. trichocarpa* metabolic pathways, which we then analyzed for higher order structures.

We produced a network of pathway-pathway associations between poplarCyc v. 3 metabolic pathways (Zhang et al., 2010) by placing edges between pathways that share a significant number of predicted PPIs between the proteins underlying the pathway, excluding predicted self-interactions. Pathway linkages are considered significant if the number of inter-pathway interactions exceeds the number found in 99.9% of randomized networks. Using the ENTS 0.65 *P. trichocarpa* network, we find 913 significant pathway linkages (Fig. 5, Additional File 2). All but 2 pairs of pathway linkages are joined within the largest connected component of the linkage graph, and the groups not connected to the primary component include one pair of sulfate metabolic pathways and one pair of heavy metal transporters. Out of the 913 total linkages, we find 173 that share at least 1 compound, significantly more than expected by chance (85.95 ± 7.90, p < 0.0001). Seven out of the 10 most highly connected pathways are involved in the biosynthesis of carbohydrates, with the sucrose biosynthesis pathway having the highest degree with 38 pathway linkages. The major entry point of reduced nitrogen, the glutamine biosynthesis pathway, is the fifth most highly connected pathway with 27 linkages. The remaining 2 most highly connected pathways include the flavonoid biosynthesis pathway and glycolysis.

We then used the MCL algorithm to produce clusters of pathways, which we assessed for enrichment of specific pathway classes (Table 2, Additional File 3). We find 9 out of 22 clusters that are significantly enriched for specific classes at a family-wise error rate of 0.05, following a Bonferroni correction. The largest cluster is highly enriched for the biosynthesis of
phenylpropanoid derivatives ($p=4.105\times10^{-11}$), which include a large variety of secondary metabolites important for structure, defense against pathogens, and defense from herbivory. A second cluster is highly enriched for the biosynthesis of fatty acids and lipids ($p = 2.838\times10^{-10}$) (Fig. 5). The fatty acid biosynthesis cluster includes not only pathways for the production of phospholipids - primarily phosphatidylcholine - but also biosynthetic pathways for hydrophobic electron carriers such as quinones and quinols. This cluster also contains the rubisco shunt, which acts as a bypass to the Calvin cycle in order to decrease carbon loss during carbohydrate to lipid conversion by approximately 40% (Schwender et al., 2004). The other pathways in this cluster include the two glycolysis pathways, which act as the other main sources of pyruvate prior to fatty acid synthesis. Two additional clusters are highly enriched for nucleoside/nucleotide biosynthesis ($p = 1.831\times10^{-13}$) and nucleoside/nucleotide degradation ($p = 2.283\times10^{-21}$), respectively (Fig. 5). Interestingly, while the degradation cluster only contains purine and pyrimidine degradation pathways and a single pyrimidine salvage pathway, the biosynthesis cluster also contains arginine biosynthesis pathways. This non-intuitive grouping is supported by work in *A. thaliana* that demonstrates the coordination of arginine biosynthesis with the biosynthesis of pyrimidines (Slocum, 2005). Amino acid biosynthesis pathways are divided over several clusters and therefore do not consistently show up as enriched within their clusters. However, the groupings of several amino acid biosynthetic pathways do reflect their biochemical commonalities. The single cluster with significant enrichment of amino acid biosynthesis groups two of the three amino acids derived from 3-phosphoglycerate - serine and cysteine - together with threonine, the only other amino acid besides serine to carry a hydroxyl group. The 3 branched chain amino acid biosynthetic pathways are also grouped together within cluster 9, while the two proline biosynthetic pathways are grouped with glutamine and glutamate.
biosynthesis in clusters 2 and 4, reflecting their common origins from α-ketoglutarate (Table 2, Additional File 3).

**Predictions of human disease associations**

The analysis of PPI networks has great potential for aiding our understanding of heritable disease, as the manifestation of a given pathology may result from the perturbation of entire network modules rather than the abrogation of a single gene (Vidal et al., 2011). In particular, the physical associations between disease-related genes within a protein interaction network may signify a functional relationship between the corresponding disease states, including co-morbidity or alternative routes to a disease due to disruption of a shared pathway.

As a demonstration of the potential for ENTS predictions to provide insights into human diseases, we created a network of associations between human diseases found in the OMIM database. Edges are inferred between diseases if the corresponding disease genes are predicted to produce interacting proteins within the ENTS 0.65 confidence human PPI network and if these interactions are more frequent than expected by random chance (see methods). This leads to 552 disease associations covering 408 distinct pathologies and divided into 61 connected components (Additional File 4). Overall, we find the disease network to be significantly enriched for similarity in the literature relative to random networks (Kolmogorov-Smirnov one-sided test; \( D=0.0948; p = 5.065\times10^{-5} \) (Fig. 6a). The network contains a variety of intuitive and non-intuitive relationships between pathologies, many of which are based on interactions absent from the public databases.

As an example of intuitive relationships, one of the connected components consists entirely of associations between 11 cataract disorders (Fig. 6b), all of which are based on predicted
interactions between crystallin proteins without experimental support in public databases. We also find a number of non-intuitive relationships with anecdotal support in the literature. These include associations within one connected component containing several myopathies, including cardiomyopathy, limb-girdle muscular dystrophy, myotilinopathy, and spheroid body myopathy. These three latter disorders have overlapping symptoms and are known to co-occur with cardiomyopathy (Olive et al., 2005), as indicated by the network (Fig. 6c). A disease characterized by progressive extraocular muscle weakness - horizontal gaze palsy with progressive scoliosis - is also associated with both cardiomyopathy and limb-girdle muscular dystrophy based on a predicted interaction between TCAP and ROBO3, though limb-girdle muscle weakness is not directly associated with this disorder in the literature (Crisfield, 1974). The last condition within the component, vesicoureteral reflux (VUR), is characterized by developmental abnormalities of the kidney and urinary tract. Its relationship to cardiomyopathy and limb-girdle muscular dystrophy is indicated by a predicted interaction between TCAP and ROBO2. Interestingly, although VUR is not associated with disorders of the striated or cardiac muscle, it has been observed to co-occur with visceral myopathy in cases of Chronic Intestinal Pseudo Obstruction and Berdon Syndrome (Andres et al., 2010).

Network properties and duplicate gene evolution

Recently, several authors have proposed a relationship between the properties of biological networks and the evolution of duplicate genes. Studies of paleopolyploid plants have demonstrated that functional categories generally associated with higher network connectivity tend to be retained in duplicate following whole genome duplication (WGD) (Rodgers-Melnick et al., 2012). These observations led to the development of the gene balance hypothesis, which predicts that more highly connected genes should tend to be retained following WGD because of
purifying selection for stoichiometric balance among interaction proteins (Birchler and Veitia, 2010). Unfortunately, the lack of large-scale PPI data for paleopolyploid plants has largely precluded a thorough network-based analysis of this phenomenon. Here, we used the predicted A. thaliana and P. trichocarpa 0.65 confidence PPI networks to conduct a preliminary analysis of the relationship between WGD duplicate retention and 2 properties of the predicted network: the fraction of genes to which a given gene is connected (degree centrality) and the fraction of neighbors retained following the same WGD (duplicated neighbors). Based on logistic regression, the fraction of duplicated neighbors and the interaction term with degree centrality was positively associated with the presence of a duplicate paralog (Fig. 7, Table 3). These results were highly consistent with those generated when we restricted the analysis to genes with at least 10 neighbors (Additional file1; Supplemental Table 2). Therefore, the duplication state of a given gene's neighbors has a strong effect on its probability of retention, and this effect is enhanced with a higher number of interactions. This fits the predictions of the gene balance hypothesis in that the dependence on connectedness strongly depends upon the dosage of the interacting genes. Interestingly, the degree centrality main effect is negatively associated with the odds of retention once the interaction term is taken into account, although this effect is not significant for the A. thaliana β/γ WGD and inconsistently significant for the α WGD.
Table 3.1 Protein-protein interactions predicted using ENTS at several confidence cutoffs. The number of genes involved in the predicted interactions is shown in parentheses.

<table>
<thead>
<tr>
<th>Training</th>
<th>Prediction</th>
<th>0.55</th>
<th>0.65</th>
<th>0.75</th>
<th>0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td><em>S. cerevisiae</em></td>
<td>29,616 (4,320)</td>
<td>10,933 (3,314)</td>
<td>2,841 (1,713)</td>
<td>597 (576)</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td><em>H. sapiens</em></td>
<td>212,365 (12,936)</td>
<td>94,082 (9,906)</td>
<td>29,562 (6,377)</td>
<td>4,180 (2,223)</td>
</tr>
<tr>
<td></td>
<td><em>M. musculus</em></td>
<td>244,548 (13,615)</td>
<td>98,108 (10,157)</td>
<td>26,860 (5,889)</td>
<td>2,825 (1,496)</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td><em>A. thaliana</em></td>
<td>346,020 (15,964)</td>
<td>176,600 (13,426)</td>
<td>79,796 (9,504)</td>
<td>19,915 (4,010)</td>
</tr>
<tr>
<td></td>
<td><em>P. trichocarpa</em></td>
<td>481,253 (19,321)</td>
<td>178,232 (14,536)</td>
<td>42,503 (7,501)</td>
<td>4,085 (1,316)</td>
</tr>
</tbody>
</table>
Table 3.2 Classes of metabolic pathway that were significantly enriched in clusters at a 0.05 family-wide type I error rate. The number of pathways present in each cluster is given in parentheses in the first column, while the number of the given class of pathway within each cluster is given by parentheses in the second column.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Class</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (31)</td>
<td>Phenylpropanoid Derivatives Biosynthesis (6)</td>
<td>4.105×10^{-11}</td>
</tr>
<tr>
<td></td>
<td>Secondary Metabolites Biosynthesis (8)</td>
<td>5.815×10^{-9}</td>
</tr>
<tr>
<td></td>
<td>Flavonoids Biosynthesis (6)</td>
<td>6.486×10^{-9}</td>
</tr>
<tr>
<td></td>
<td>Carbohydrates Biosynthesis (9)</td>
<td>2.185×10^{-8}</td>
</tr>
<tr>
<td></td>
<td>Sugars Biosynthesis (8)</td>
<td>5.65×10^{-6}</td>
</tr>
<tr>
<td>2 (22)</td>
<td>Nitrogen Compounds Metabolism (3)</td>
<td>1.077×10^{-5}</td>
</tr>
<tr>
<td></td>
<td>Inorganic Nutrients Metabolism (3)</td>
<td>1.134×10^{-5}</td>
</tr>
<tr>
<td>3 (21)</td>
<td>Fatty Acids and Lipids Biosynthesis (8)</td>
<td>2.838×10^{-10}</td>
</tr>
<tr>
<td></td>
<td>Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis (8)</td>
<td>1.741×10^{-7}</td>
</tr>
<tr>
<td></td>
<td>Phospholipid Biosynthesis (6)</td>
<td>2.106×10^{-7}</td>
</tr>
<tr>
<td></td>
<td>Quinol and Quinone Biosynthesis (3)</td>
<td>6.301×10^{-6}</td>
</tr>
<tr>
<td>4 (17)</td>
<td>Vitamins Biosynthesis (4)</td>
<td>3.871×10^{-6}</td>
</tr>
<tr>
<td>8 (12)</td>
<td>Nucleosides and Nucleotides Biosynthesis (6)</td>
<td>1.831×10^{-13}</td>
</tr>
<tr>
<td></td>
<td>Purine Nucleotide Biosynthesis (3)</td>
<td>7.213×10^{-6}</td>
</tr>
<tr>
<td>11 (6)</td>
<td>Nucleosides and Nucleotides Degradation (5)</td>
<td>2.283×10^{-21}</td>
</tr>
<tr>
<td></td>
<td>Purine Nucleotides Degradation (4)</td>
<td>2.128×10^{-13}</td>
</tr>
<tr>
<td></td>
<td>Degradation/Utilization/Assimilation (5)</td>
<td>3.977×10^{-10}</td>
</tr>
<tr>
<td>15 (3)</td>
<td>Amino Acids Biosynthesis (3)</td>
<td>1.642×10^{-6}</td>
</tr>
<tr>
<td>22 (2)</td>
<td>Inorganic Nutrients Metabolism (2)</td>
<td>2.583×10^{-7}</td>
</tr>
<tr>
<td></td>
<td>Sulfur Compounds Metabolism (2)</td>
<td>1.062×10^{-6}</td>
</tr>
<tr>
<td>23 (2)</td>
<td>Transport (2)</td>
<td>1.256×10^{-3}</td>
</tr>
</tbody>
</table>
Table 3.3 Estimated coefficients and significance for the logistic regressions on WGD retention in *P. trichocarpa* and *A. thaliana* using network properties degree centrality (DC) and duplicated neighbors (DN) from the ENTS 0.65 confidence networks. P-values are based on confidence intervals generated from fitting the model on resampled data 10,000 times.

<table>
<thead>
<tr>
<th>WGD</th>
<th>Intercept</th>
<th>log(DC)</th>
<th>DN</th>
<th>log(DC):DN</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. trichocarpa</em> Salicoid</td>
<td>-1.90193***</td>
<td>-0.24628***</td>
<td>4.91163***</td>
<td>0.57953***</td>
</tr>
<tr>
<td><em>A. thaliana</em> α</td>
<td>-1.98648***</td>
<td>-0.09509***</td>
<td>5.73056***</td>
<td>0.69114***</td>
</tr>
<tr>
<td><em>A. thaliana</em> β/γ</td>
<td>-2.49904***</td>
<td>-0.03628***</td>
<td>6.93515***</td>
<td>0.72433***</td>
</tr>
</tbody>
</table>

*p<0.05; **p<0.01; ***p<0.001
Figure 3.1 The ROC curves for ENTS testing data on (A) *A. thaliana*, (B) *H. sapiens*, and (C) *S. cerevisiae*. 
Figure 3.2 The number of positive predictions on sets of novel experimentally-determined interactions for ENTS and several alternative prediction techniques in (A) *A. thaliana*, (B) *S. cerevisiae*, (C) *H. sapiens*, and (D) *M. musculus*. ENTS results are shown in blue; the alternative classifiers’ results are shown in green; and the number of positive predictions shared between each pair of classifiers is shown in yellow.
Figure 3.3 (A) The frequency of shared KEGG pathways and (B) the mean GO molecular function similarity scores for predicted ENTS interactions. Vertical lines at each confidence level show the total range for randomized networks. (C) Shared KEGG pathway frequency and (D) mean GO molecular function similarity for ENTS as compared to other predicted networks and the experimentally-verified network. The experimentally-verified networks are shown on the left for each organism, with a hatched bar.
Figure 3.4 Dots show the observed frequency of interactors with $\rho > 0.5$. Error bars indicate 95% confidence intervals based on 250 bootstrapped replicates. Grey bars indicated 95% confidence intervals from 250 randomized networks. The leftmost lane for each organism with the exception of P. trichocarpa is the data for the experimentally-verified network.
Figure 3.5 Each node represents a pathway in the PoplarCyc metabolic network, with lines connecting those with a significant number of interpathway interactions based on ENTS predictions. Red lines indicate a shared compound between pathways. The yellow nodes belong to a cluster significantly enriched for fatty acid biosynthesis, while the green and blue nodes belong to pathways significantly enriched for nucleotide/nucleoside biosynthesis and degradation, respectively.
Figure 3.6 (A) The cumulative distribution of literature similarity scores in randomized networks (green) and the observed human disease network constructed from ENTS PPI predictions (blue). (B) and (C) connected components in the human disease network. Line widths are drawn proportional to literature similarity scores.
Figure 3.7 The relationship between degree centrality, the fraction of predicted neighbors retained following WGD, and the probability of whether a gene in the predicted network is retained following the WGD for the (A) *P. trichocarpa* Salicoid duplication, (B) the *A. thaliana* α duplication, and (C) the *A. thaliana* β and γ duplications.
Discussion

ENTS performance

In this study we find that our random forest-based classifier, ENTS, can predict biologically meaningful PPIs both within the species on which we trained the classifier and within species sharing a relatively distant common ancestor with the training species. These results are comparable or favorable to existing methods of protein interaction prediction, including several that used experimentally-obtained functional data as predictors. This demonstrates that researchers may generate a high quality set of probable PPIs prior to performing extensive functional experimentation. Therefore these predicted PPIs may serve as a basis for the development of functional hypotheses in newly-sequence organisms. Notably, even in the cases in which ENTS predictions were outperformed - the *A. thaliana* PAIR predictor and the *M. musculus* Yellaboina interolog approach - the majority of interactions predicted by ENTS were not predicted by the alternative approach. This suggests the value of using ENTS as part of an ensemble rather than relying on any single classifier. This may be particularly effective when large amounts of functional data are available, as with PAIR, or an orthologous interactome is well-covered, as with the Yellaboina interolog approach.

The performance of ENTS does vary significantly between organisms, with the AUC ranging from a low of 0.811 in *S. cerevisiae* to a high of 0.963 in *A. thaliana*. These differences likely arise due to functional differences between the positive datasets. The contrast in performance between the *A. thaliana* and non-plant classifiers may be attributed to biases in the interaction data currently available for these organisms - particularly *A. thaliana*. With the
exception of a recent high-throughput study that was not used for training (Consortium, 2011). *A. thaliana* studies of PPIs within the literature have focused on testing of specific hypotheses concerning proteins of high *a priori* importance. This can result in sets of highly clustered, high degree nodes within literature-curated PPI networks (Consortium, 2011; Cusick et al., 2009). Such clusters have likely led to an enrichment of *A. thaliana* interacting pairs with high pairwise domain odds (Additional file 1; Supplemental Fig. 1). In contrast, the multiple high throughput studies of binary PPIs conducted on *H. sapiens* (Stelzl et al., 2005), *S. cerevisiae* (Yu et al., 2008), and *D. melanogaster* (Guruharsha et al., 2011) contribute a more unbiased view of the interactome than for *A. thaliana*. The more comprehensive experimental datasets have yielded a number of interactions that are less amenable to ENTS detection due to either an absence of PFAM domains in one or both of the proteins in the pair or because of the presence of domain pairs infrequently associated with physical interaction. Testing performance based primarily on literature-curated sets may therefore give results that are inconsistent with the true global protein interaction network. This highlights a necessary caveat present for all statistical learning methods - the performance of the classifier on novel data depends on the scope of its training).

Although the most important features for ENTS classification are based on domain composition, we also find subcellular localization information from MultiLoc2 to be highly important for prediction of PPIs. Interestingly, the most important MultiLoc2 features tend to be the SVM amino acid scores, which indicate how well the total amino acid content of a protein matches a particular subcellular compartment (Blum et al., 2009). These measures provided ENTS with more sensitivity to subtle changes in amino acid content than the domain-based scores, which rely on the gain or loss of conserved PFAM domains. For instance, the duplicated genes from the Salicoid WGD in *P. trichocarpa* only share 56% of their predicted neighbors on
average in the 0.65 confidence network despite high pairwise similarity in domain content. This allows for the possibility that ENTS may predict rewiring of PPIs within networks following duplication, although confirmation of this would require experimental validation.

**Interpretation of ENTS interactions**

We trained ENTS using sets of known physically-interacting pairs of proteins. Therefore, proteins predicted to interact by ENTS should be interpreted as having a high potential for physical interaction, given that they are present within the same location at the same time. Each set of predicted interactions represents a more than 1000-fold reduction in the number of total possible interactions for the organism. However, because we want to permit the use of this classifier across a variety of organisms without broad functional data, we do not include expression or proteomics data as features in the set of predictor variables. As such, researchers should confirm all predictions with functional data. Furthermore, ENTS has limited capability for predicting interactions between proteins that either lack conserved domains or that contain domains never before experimentally observed within physically interacting proteins. This precludes the detection of some novel interactions, although other researchers may wish to append functional data to the set of ENTS predictor variables for their organism in order to predict these interactions through a greater variety of evidence sources.

**Applications of ENTS networks**

The elucidation of PPI networks permits tremendous insight into both cellular and evolutionary processes. In humans, several studies have used experimental PPI data to map relationships between human diseases based on causative similarity and infer causative disease genes within large implicated linkage regions (Lage et al., 2007). However, such studies have
reported limitations due to the lack of detected interactions with shared disease etiology. Here, we have shown that ENTS is capable of yielding novel predicted interactions with relevance to known co-occurring human diseases. We contend that these predicted interactions may be used by biomedical researchers to narrow the scope of regions implicated in genome-wide association studies in addition to providing predictions on which to base more targeted searches for candidate loci. The latter function may prove especially vital due to the increasingly visible role of rare de novo mutations, particularly CNVs, in the etiology of disease (Malhotra and Sebat, 2012). Furthermore, the generality of our approach across species allows the potential for yielding insight into a variety of agriculturally important diseases in non-model species that are currently only understood through large linkage regions based on QTL and relatively small GWAS studies. We have also shown that ENTS is capable of revealing the higher-order structure of metabolic networks in a plant species without extensive experimental data, *P. trichocarpa*. ENTS-predicted physical interactions yield a significant enrichment of associations between pathways that share compounds and group sets of coordinated pathways such as pyrimidine and arginine biosynthesis (Slocum, 2005). Such insights may be used to inform targets of selection in breeding programs in order to increase the output flux from key pathways.

We have demonstrated that ENTS has great potential for yielding insight into network evolution in non-model species across the plant, animal, and fungal kingdoms. Networks have come to play an increasingly central role in evolutionary studies (De Smet and Van de Peer, 2012), but a rigorous analysis of network evolution requires the development of whole-genome networks for many non-model species. Here, we have shown that the topological properties of ENTS-predicted networks, including the first whole-genome interactome for *P. trichocarpa*, are related to the probability of WGD duplicate gene retention following independent duplication
events in a manner consistent with the predictions of the gene balance hypothesis. Specifically, we have shown that the duplication state of a gene's neighbors has a strong impact on the probability of retention following WGD and that this effect is enhanced at higher connectivity, which is consistent with the hypothesis of a selective drive to maintain stoichiometric balance between interacting proteins. Interestingly, the degree centrality main effect was only significant for the *P. trichocarpa* Salicoid WGD and the *A. thaliana* α WGD. In the case of *P. trichocarpa*, the negative degree centrality coefficient leads to an inverse relationship between connectivity and the probability of duplicate retention in the presence of few duplicated neighbors. This decrease is, again, consistent with the gene balance hypothesis, as the effects of stoichiometric imbalance are likely to become more extreme at higher connectivity and would therefore favor a singleton state when interconnected genes are also singletons. Intriguingly, this effect is also present in the *A. thaliana* duplications, though it is highly diminished with respect to *P. trichocarpa* and shows inconsistent statistical significance. One explanation is that *A. thaliana* has undergone a greater degree of fractionation following its last two WGDs than has *P. trichocarpa* following the Salicoid WGD (Van de Peer et al., 2009), so the influence of connectivity is more apparent in the latter. This suggests that the *P. trichocarpa* duplicate genes most at risk for future nonfunctionalization include those with low degree centrality and few duplicated neighbors.

**Conclusions**

We have introduced an efficient new approach that enables prediction of protein-protein interactions on a whole genome scale based entirely on information that can be derived from primary sequence data. This is a potentially groundbreaking addition to the standard toolbox for newly-sequence non-model genomes, which are rapidly proliferating. The networks derived from
our protein-protein interaction predictions are realistic from the standpoint of consistency with co-expression and shared functional annotations of connected genes. Furthermore, we have shown that our predictions can reveal supported relationships among emergent phenotypes such as human disease states and the coordination of metabolic pathways. Finally, we have demonstrated that our inferred networks can reveal subtle details of genome-scale evolution. Because the method can readily be applied on a large scale to phylogenetically-diverse organisms, we anticipate that large-scale comparative analyses will provide insights into the mechanisms of network structure evolution.
Acknowledgements

This work was supported 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, and through a United States Department of Agriculture National Institute of Food and Agriculture pre-doctoral fellowship.
References


121


Supplemental Table 3.1 Predictors used for the ENTS classifier. Importance measures indicate the mean decrease in accuracy in the out-of-bag training data upon permutation of that variable.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>A. thaliana Importance</th>
<th>S. cerevisae Importance</th>
<th>H. sapiens Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum_odds</td>
<td>Sum of the odds scores for unique pairs of domains in the protein pair</td>
<td>0.0621</td>
<td>0.0536</td>
<td>0.0521</td>
</tr>
<tr>
<td>highest_odds</td>
<td>The highest observed odds score for a pair of domains in the protein pair</td>
<td>0.0604</td>
<td>0.0530</td>
<td>0.0578</td>
</tr>
<tr>
<td>lowest_odds</td>
<td>The lowest observed odds score for a pair of domains in the protein pair</td>
<td>0.0500</td>
<td>0.0414</td>
<td>0.0313</td>
</tr>
<tr>
<td>not_observed</td>
<td>The number of unique domain pairs in the protein pair never observed in any known interactions</td>
<td>0.0573</td>
<td>0.0646</td>
<td>0.0627</td>
</tr>
<tr>
<td>not_observed_frac</td>
<td>The fraction of unique domain pairs in the protein pair never observed in any known interactions</td>
<td>0.0594</td>
<td>0.0470</td>
<td>0.0564</td>
</tr>
<tr>
<td>n_odds_pairs</td>
<td>The number of unique domain pairs observed in at least 1 known interaction</td>
<td>0.0441</td>
<td>0.0404</td>
<td>0.0435</td>
</tr>
<tr>
<td>n_domine_pairs</td>
<td>The number of unique domain pairs predicted or verified to interact in the DOMINE database</td>
<td>0.0370</td>
<td>0.0247</td>
<td>0.0342</td>
</tr>
<tr>
<td>highest_domine_conf</td>
<td>The highest confidence score for a pair of domains verified to interact in the DOMINE database</td>
<td>0.0432</td>
<td>0.0273</td>
<td>0.0389</td>
</tr>
<tr>
<td>cytoplasmic1</td>
<td>The MultiLoc2 probability of localization in the cytoplasm for the first protein in the pair</td>
<td>0.0313</td>
<td>0.0458</td>
<td>0.0445</td>
</tr>
<tr>
<td>nuclear1</td>
<td>The MultiLoc2 probability of localization in the nucleus for the first protein in the pair</td>
<td>0.0420</td>
<td>0.0480</td>
<td>0.0394</td>
</tr>
<tr>
<td>peroxisomal1</td>
<td>The MultiLoc2 probability of localization in the peroxisome for the first protein in the pair</td>
<td>0.0421</td>
<td>0.0367</td>
<td>0.0383</td>
</tr>
<tr>
<td>ER1</td>
<td>The MultiLoc2 probability of localization in the ER for the first protein in the pair</td>
<td>0.0305</td>
<td>0.0289</td>
<td>0.0336</td>
</tr>
<tr>
<td>mitochondrial1</td>
<td>The MultiLoc2 probability of localization in the mitochondrion for the first protein in the pair</td>
<td>0.0384</td>
<td>0.0352</td>
<td>0.0423</td>
</tr>
<tr>
<td>extracellular1</td>
<td>The MultiLoc2 probability of localization in the extracellular space for the first protein in the pair</td>
<td>0.0371</td>
<td>0.0278</td>
<td>0.0316</td>
</tr>
<tr>
<td>plasma membrane1</td>
<td>The MultiLoc2 probability of localization in the plasma membrane for the first protein in the pair</td>
<td>0.0306</td>
<td>0.0301</td>
<td>0.0326</td>
</tr>
<tr>
<td>Golgi apparatus1</td>
<td>The MultiLoc2 probability of localization in the Golgi apparatus for the first protein in the pair</td>
<td>0.0250</td>
<td>0.0355</td>
<td>0.0354</td>
</tr>
<tr>
<td>chloroplast1</td>
<td>The MultiLoc2 probability of localization in the chloroplast for the first protein in the pair (plant only)</td>
<td>0.0193</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>vacuolar1</td>
<td>The MultiLoc2 probability of localization in the vacuole for the first protein in the pair (plant and fungal)</td>
<td>0.0316</td>
<td>0.0237</td>
<td>N/A</td>
</tr>
<tr>
<td>lysosomal1</td>
<td>The MultiLoc2 probability of localization in the lysosome for the first protein in the pair (animal only)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0318</td>
</tr>
<tr>
<td>svm_target_score_sp1</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for secretory pathways in the first protein of the pair</td>
<td>0.0439</td>
<td>0.0292</td>
<td>0.0433</td>
</tr>
<tr>
<td>svm_target_score_mtp1</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for mitochondrion in the first protein of the pair</td>
<td>0.0300</td>
<td>0.0351</td>
<td>0.0390</td>
</tr>
<tr>
<td>svm_target_scorectp1</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for chloroplast in the first protein of the pair (plant only)</td>
<td>0.0183</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>svm_target_score_mtp_vsctp1</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides contrasting mitochondrion with chloroplast</td>
<td>0.0395</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>svm_target_score_oth1</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for other compartments in the first protein of the pair</td>
<td>0.0252</td>
<td>0.0279</td>
<td>0.0367</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>svm_sa_score1</td>
<td>MultiLoc2 score for presence of signal anchor for membrane proteins of secretory pathway in the first protein of the pair</td>
<td>0.0438</td>
<td>0.0470</td>
<td>0.0384</td>
</tr>
<tr>
<td>svm_aac_score_nuc1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the nucleus</td>
<td>0.0313</td>
<td>0.0464</td>
<td>0.0382</td>
</tr>
<tr>
<td>svm_aac_score_cyt1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the cytoplasm</td>
<td>0.0409</td>
<td>0.0481</td>
<td>0.0497</td>
</tr>
<tr>
<td>svm_aac_score_nuc_vs_cyt1</td>
<td>MultiLoc2 score for contrasting the overall amino acid content of the first protein with respect to nuclear vs. cytoplasm localization</td>
<td>0.0234</td>
<td>0.0421</td>
<td>0.0318</td>
</tr>
<tr>
<td>svm_aac_score_mit1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the mitochondrion</td>
<td>0.0474</td>
<td>0.0474</td>
<td>0.0501</td>
</tr>
<tr>
<td>svm_aac_score_chl1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein (plant) matches the chloroplast</td>
<td>0.0504</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>svm_aac_score_ext1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the extracellular space</td>
<td>0.0503</td>
<td>0.0490</td>
<td>0.0503</td>
</tr>
<tr>
<td>svm_aac_score_pm1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the plasma membrane</td>
<td>0.0508</td>
<td>0.0525</td>
<td>0.0544</td>
</tr>
<tr>
<td>svm_aac_score_per1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the peroxisome</td>
<td>0.0420</td>
<td>0.0460</td>
<td>0.0522</td>
</tr>
<tr>
<td>svm_aac_score_er1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the ER</td>
<td>0.0511</td>
<td>0.0409</td>
<td>0.0506</td>
</tr>
<tr>
<td>svm_aac_score_gol1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the Golgi apparatus</td>
<td>0.0501</td>
<td>0.0512</td>
<td>0.0543</td>
</tr>
<tr>
<td>svm_aac_score_vac1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the vacuole (plant and fungal)</td>
<td>0.0441</td>
<td>0.0442</td>
<td>N/A</td>
</tr>
<tr>
<td>goloc_score_nuc1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the nucleus</td>
<td>0.0406</td>
<td>0.0344</td>
<td>0.0462</td>
</tr>
<tr>
<td>goloc_score_cyt1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the cytoplasm</td>
<td>0.0370</td>
<td>0.0364</td>
<td>0.0397</td>
</tr>
<tr>
<td>goloc_score_mit1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the mitochondrion</td>
<td>0.0381</td>
<td>0.0342</td>
<td>0.0404</td>
</tr>
<tr>
<td>goloc_score_chl1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the chloroplast (plant)</td>
<td>0.0277</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>goloc_score_ext1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the extracellular space</td>
<td>0.0307</td>
<td>0.0287</td>
<td>0.0470</td>
</tr>
<tr>
<td>goloc_score_pm1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the plasma membrane</td>
<td>0.0356</td>
<td>0.0299</td>
<td>0.0441</td>
</tr>
<tr>
<td>goloc_score_per1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the peroxisome</td>
<td>0.0433</td>
<td>0.0335</td>
<td>0.0387</td>
</tr>
<tr>
<td>goloc_score_er1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the ER</td>
<td>0.0375</td>
<td>0.0312</td>
<td>0.0466</td>
</tr>
<tr>
<td>goloc_score_gol1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the Golgi apparatus</td>
<td>0.0310</td>
<td>0.0333</td>
<td>0.0405</td>
</tr>
<tr>
<td>goloc_score_vac1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the vacuole (plant and fungus)</td>
<td>0.0400</td>
<td>0.0372</td>
<td>N/A</td>
</tr>
<tr>
<td>phyloloc_score_nuc1</td>
<td>MultiLoc2 score for whether phylogenetic profile of</td>
<td>0.0426</td>
<td>0.0433</td>
<td>0.0484</td>
</tr>
<tr>
<td>Protein ID/Score</td>
<td>Description</td>
<td>First BLAST Hit Score</td>
<td>Second BLAST Hit Score</td>
<td>Combined Score</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------------------</td>
<td>------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>phyloc_score_cyt1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates cytoplasmic localization</td>
<td>0.0458</td>
<td>0.0468</td>
<td>0.0476</td>
</tr>
<tr>
<td>phyloc_score_mit1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates mitochondrial localization</td>
<td>0.0470</td>
<td>0.0560</td>
<td>0.0494</td>
</tr>
<tr>
<td>phyloc_score_chl1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates chloroplast localization (plant)</td>
<td>0.0471</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>phyloc_score_ext1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates extracellular localization</td>
<td>0.0521</td>
<td>0.0468</td>
<td>0.0468</td>
</tr>
<tr>
<td>phyloc_score_pm1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates plasma membrane localization</td>
<td>0.0485</td>
<td>0.0508</td>
<td>0.0501</td>
</tr>
<tr>
<td>phyloc_score_per1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates peroxisomal localization</td>
<td>0.0512</td>
<td>0.0533</td>
<td>0.0498</td>
</tr>
<tr>
<td>phyloc_score_er1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates ER localization</td>
<td>0.0503</td>
<td>0.0537</td>
<td>0.0517</td>
</tr>
<tr>
<td>phyloc_score_gol1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates Golgi apparatus localization</td>
<td>0.0446</td>
<td>0.0475</td>
<td>0.0511</td>
</tr>
<tr>
<td>phyloc_score_vac1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates vacuolar localization (plant and fungal)</td>
<td>0.0507</td>
<td>0.0529</td>
<td>N/A</td>
</tr>
<tr>
<td>motif_search_er_target1</td>
<td>MultiLoc2 indicator of ER targeting signal for the first protein</td>
<td>0.0019</td>
<td>0.0011</td>
<td>0.0013</td>
</tr>
<tr>
<td>motif_search_pm_receptor_domain1</td>
<td>MultiLoc2 indicator of plasma membrane targeting signal for the first protein</td>
<td>0.0000</td>
<td>-0.0079</td>
<td>0.0166</td>
</tr>
<tr>
<td>motif_search_peroxisomal_target1</td>
<td>MultiLoc2 indicator of peroxisomal targeting signal for the first protein</td>
<td>0.0092</td>
<td>-0.0032</td>
<td>0.0101</td>
</tr>
<tr>
<td>motif_search_nuclear_bipartite1</td>
<td>MultiLoc2 indicator of nuclear bitartite protein for the first protein</td>
<td>0.0118</td>
<td>0.0111</td>
<td>0.0111</td>
</tr>
<tr>
<td>motif_search_predict-NLS1</td>
<td>MultiLoc2 indicator of nuclear localization signal for the first protein</td>
<td>0.0144</td>
<td>0.0138</td>
<td>0.0150</td>
</tr>
<tr>
<td>motif_search_nls_mon1</td>
<td>MultiLoc2 indicator of nuclear localization signal for the first protein</td>
<td>0.0160</td>
<td>0.0219</td>
<td>0.0243</td>
</tr>
<tr>
<td>motif_search_dna_associated_domain1</td>
<td>MultiLoc2 indicator of dna associated domain for the first protein</td>
<td>0.0183</td>
<td>0.0062</td>
<td>0.0253</td>
</tr>
<tr>
<td>phyloc_score_lys1</td>
<td>MultiLoc2 score for whether phylogenetic profile of first protein BLAST hits indicates lysosome localization (animal)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0547</td>
</tr>
<tr>
<td>goloc_score_lys1</td>
<td>MultiLoc2 score for whether inferred GO categories of the first protein match the lysosome membrane (animal)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0422</td>
</tr>
<tr>
<td>svm_aac_score_lys1</td>
<td>MultiLoc2 score for whether the overall amino acid content of the first protein matches the lysosome (animal)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0507</td>
</tr>
<tr>
<td>cytoplasmic2</td>
<td>The MultiLoc2 probability of localization in the cytoplasm for the first protein in the pair</td>
<td>0.0383</td>
<td>0.0410</td>
<td>0.0403</td>
</tr>
<tr>
<td>nuclear2</td>
<td>The MultiLoc2 probability of localization in the nucleus for the second protein in the pair</td>
<td>0.0455</td>
<td>0.0486</td>
<td>0.0364</td>
</tr>
<tr>
<td>peroxisomal2</td>
<td>The MultiLoc2 probability of localization in the peroxisome for the second protein in the pair</td>
<td>0.0418</td>
<td>0.0379</td>
<td>0.0442</td>
</tr>
<tr>
<td>ER2</td>
<td>The MultiLoc2 probability of localization in the ER for</td>
<td>0.0323</td>
<td>0.0211</td>
<td>0.0276</td>
</tr>
<tr>
<td>Location</td>
<td>Description</td>
<td>SVM MultiLoc2 Probability</td>
<td>SVM Target Score Sp</td>
<td>SVM Target Score Mtp</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>---------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>mitochondrial2</td>
<td>The MultiLoc2 probability of localization in the mitochondrial for the second protein in the pair</td>
<td>0.0315</td>
<td>0.0403</td>
<td>0.0312</td>
</tr>
<tr>
<td>extracellular2</td>
<td>The MultiLoc2 probability of localization in the extracellular space for the second protein in the pair</td>
<td>0.0372</td>
<td>0.0323</td>
<td>0.0344</td>
</tr>
<tr>
<td>plasma membrane2</td>
<td>The MultiLoc2 probability of localization in the plasma membrane for the second protein in the pair</td>
<td>0.0352</td>
<td>0.0329</td>
<td>0.0371</td>
</tr>
<tr>
<td>Golgi apparatus2</td>
<td>The MultiLoc2 probability of localization in the Golgi apparatus for the second protein in the pair</td>
<td>0.0307</td>
<td>0.0364</td>
<td>0.0357</td>
</tr>
<tr>
<td>chloroplast2</td>
<td>The MultiLoc2 probability of localization in the chloroplast for the second protein in the pair (plant only)</td>
<td>0.0255</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>vacuolar2</td>
<td>The MultiLoc2 probability of localization in the vacuole for the second protein in the pair (plant and fungal)</td>
<td>0.0344</td>
<td>0.0283</td>
<td>N/A</td>
</tr>
<tr>
<td>lysosomal2</td>
<td>The MultiLoc2 probability of localization in the lysosome for the second protein in the pair (animal only)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0303</td>
</tr>
<tr>
<td>svm_target_score_sp2</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for secretory pathways in the second protein of the pair</td>
<td>0.0442</td>
<td>0.0374</td>
<td>0.0431</td>
</tr>
<tr>
<td>svm_target_score_mtp2</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for mitochondrion in the second protein of the pair</td>
<td>0.0326</td>
<td>0.0366</td>
<td>0.0357</td>
</tr>
<tr>
<td>svm_target_score_ctp2</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for chloroplast in the second protein of the pair (plant only)</td>
<td>0.0314</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>svm_target_score_mtp_vs_ctp2</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides contrasting mitochondrion with chloroplast in the second protein of the pair</td>
<td>0.0442</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>svm_target_score_oth2</td>
<td>MultiLoc2 score for detection of N-terminal targeting peptides for other compartments in the second protein of the pair</td>
<td>0.0367</td>
<td>0.0375</td>
<td>0.0397</td>
</tr>
<tr>
<td>svm_sa_score2</td>
<td>MultiLoc2 score for presence of signal anchor for membrane proteins of secretory pathway in the second protein of the pair</td>
<td>0.0436</td>
<td>0.0423</td>
<td>0.0396</td>
</tr>
<tr>
<td>svm_aac_score_nuc2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the nucleus</td>
<td>0.0346</td>
<td>0.0411</td>
<td>0.0445</td>
</tr>
<tr>
<td>svm_aac_score_cyt2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the cytoplasm</td>
<td>0.0253</td>
<td>0.0504</td>
<td>0.0505</td>
</tr>
<tr>
<td>svm_aac_score_nuc_vs_cyt2</td>
<td>MultiLoc2 score for contrasting the overall amino acid content of the second protein with respect to nuclear vs. cytoplasm localization</td>
<td>0.0281</td>
<td>0.0349</td>
<td>0.0380</td>
</tr>
<tr>
<td>svm_aac_score_mit2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the mitochondrion</td>
<td>0.0513</td>
<td>0.0480</td>
<td>0.0551</td>
</tr>
<tr>
<td>svm_aac_score_chl2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein (plant) matches the chloroplast</td>
<td>0.0522</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>svm_aac_score_ext2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the extracellular space</td>
<td>0.0521</td>
<td>0.0542</td>
<td>0.0537</td>
</tr>
<tr>
<td>svm_aac_score_pm2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the plasma membrane</td>
<td>0.0508</td>
<td>0.0528</td>
<td>0.0555</td>
</tr>
<tr>
<td>svm_aac_score_per2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the peroxisome</td>
<td>0.0485</td>
<td>0.0420</td>
<td>0.0536</td>
</tr>
<tr>
<td>Function</td>
<td>Description</td>
<td>SVM Score 1</td>
<td>SVM Score 2</td>
<td>SVM Score 3</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>svm_aac_score_er2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the ER</td>
<td>0.0474</td>
<td>0.0425</td>
<td>0.0477</td>
</tr>
<tr>
<td>svm_aac_score_gol2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the Golgi apparatus</td>
<td>0.0454</td>
<td>0.0506</td>
<td>0.0519</td>
</tr>
<tr>
<td>svm_aac_score_vac2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the vacuole (plant and fungal)</td>
<td>0.0458</td>
<td>0.0434</td>
<td>N/A</td>
</tr>
<tr>
<td>goloc_score_nuc2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the nucleus</td>
<td>0.0377</td>
<td>0.0322</td>
<td>0.0440</td>
</tr>
<tr>
<td>goloc_score_cyt2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the cytoplasm</td>
<td>0.0413</td>
<td>0.0308</td>
<td>0.0337</td>
</tr>
<tr>
<td>goloc_score_mit2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the mitochondrion</td>
<td>0.0380</td>
<td>0.0328</td>
<td>0.0388</td>
</tr>
<tr>
<td>goloc_score_chl2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the chloroplast (plant)</td>
<td>0.0380</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>goloc_score_ext2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the extracellular space</td>
<td>0.0386</td>
<td>0.0283</td>
<td>0.0436</td>
</tr>
<tr>
<td>goloc_score_pm2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the plasma membrane</td>
<td>0.0392</td>
<td>0.0352</td>
<td>0.0435</td>
</tr>
<tr>
<td>goloc_score_per2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the peroxisome</td>
<td>0.0368</td>
<td>0.0357</td>
<td>0.0365</td>
</tr>
<tr>
<td>goloc_score_er2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the ER</td>
<td>0.0315</td>
<td>0.0299</td>
<td>0.0392</td>
</tr>
<tr>
<td>goloc_score_gol2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the Golgi apparatus</td>
<td>0.0279</td>
<td>0.0357</td>
<td>0.0433</td>
</tr>
<tr>
<td>goloc_score_vac2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the vacuole (plant and fungal)</td>
<td>0.0399</td>
<td>0.0330</td>
<td>N/A</td>
</tr>
<tr>
<td>phyloloc_score_nuc2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates nuclear localization</td>
<td>0.0443</td>
<td>0.0435</td>
<td>0.0476</td>
</tr>
<tr>
<td>phyloloc_score_cyt2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates cytoplasmic localization</td>
<td>0.0414</td>
<td>0.0473</td>
<td>0.0492</td>
</tr>
<tr>
<td>phyloloc_score_mit2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates mitochondrial localization</td>
<td>0.0500</td>
<td>0.0548</td>
<td>0.0512</td>
</tr>
<tr>
<td>phyloloc_score_chl2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates chloroplast localization (plant)</td>
<td>0.0451</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>phyloloc_score_ext2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates extracellular localization</td>
<td>0.0509</td>
<td>0.0475</td>
<td>0.0483</td>
</tr>
<tr>
<td>phyloloc_score_pm2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates plasma membrane localization</td>
<td>0.0497</td>
<td>0.0490</td>
<td>0.0521</td>
</tr>
<tr>
<td>phyloloc_score_per2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates peroxisomal localization</td>
<td>0.0489</td>
<td>0.0524</td>
<td>0.0491</td>
</tr>
<tr>
<td>phyloloc_score_er2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates ER localization</td>
<td>0.0498</td>
<td>0.0521</td>
<td>0.0528</td>
</tr>
<tr>
<td>phyloloc_score_gol2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates Golgi apparatus localization</td>
<td>0.0435</td>
<td>0.0474</td>
<td>0.0506</td>
</tr>
<tr>
<td>phyloloc_score_vac2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates vacuolar localization</td>
<td>0.0502</td>
<td>0.0578</td>
<td>N/A</td>
</tr>
<tr>
<td>Motif Search</td>
<td>Description</td>
<td>Score 1</td>
<td>Score 2</td>
<td>Score 3</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>motif_search_er_target2</td>
<td>MultiLoc2 indicator of ER targeting signal for the second protein</td>
<td>0.0008</td>
<td>0.0056</td>
<td>0.0030</td>
</tr>
<tr>
<td>motif_search_pm_receptor_domain2</td>
<td>MultiLoc2 indicator of plasma membrane targeting signal for the second protein</td>
<td>0.0000</td>
<td>-0.0044</td>
<td>0.0159</td>
</tr>
<tr>
<td>motif_search_peroxi_target2</td>
<td>MultiLoc2 indicator of peroxisomal targeting signal for the second protein</td>
<td>0.0083</td>
<td>0.0062</td>
<td>0.0119</td>
</tr>
<tr>
<td>motif_search_nuclear_bipartite2</td>
<td>MultiLoc2 indicator of nuclear bitartite protein for the second protein</td>
<td>0.0094</td>
<td>0.0113</td>
<td>0.0103</td>
</tr>
<tr>
<td>motif_search_predict-NLS2</td>
<td>MultiLoc2 indicator of nuclear localization signal for the second protein</td>
<td>0.0136</td>
<td>0.0111</td>
<td>0.0127</td>
</tr>
<tr>
<td>motif_search_nlsモノ2</td>
<td>MultiLoc2 indicator of nuclear localization signal for the second protein</td>
<td>0.0203</td>
<td>0.0168</td>
<td>0.0203</td>
</tr>
<tr>
<td>motif_search_dna-associated_domain2</td>
<td>MultiLoc2 indicator of dna associated domain for the second protein</td>
<td>0.0160</td>
<td>0.0019</td>
<td>0.0232</td>
</tr>
<tr>
<td>phyloloc_score_lys2</td>
<td>MultiLoc2 score for whether phylogenetic profile of second protein BLAST hits indicates lysosome localization (animal)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0534</td>
</tr>
<tr>
<td>goloc_score_lys2</td>
<td>MultiLoc2 score for whether inferred GO categories of the second protein match the lysosome membrane (animal)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0418</td>
</tr>
<tr>
<td>svm_aac_score_lys2</td>
<td>MultiLoc2 score for whether the overall amino acid content of the second protein matches the lysosome (animal)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0523</td>
</tr>
</tbody>
</table>
Supplemental Table 3.2 95% confidence intervals for coefficients in the logistic regressions on WGD retention in *P. trichocarpa* and *A. thaliana*. Models were fit using either all genes with at least one interaction in the ENTS predicted network or with the subset with at least 10 neighbors. Confidence intervals were based on non-parametric bootstrapping with 10,000 random resamplings.

<table>
<thead>
<tr>
<th>WGD</th>
<th>Data</th>
<th>Intercept</th>
<th>log(degree centrality)</th>
<th>duplicated neighbors</th>
<th>log(degree centrality):duplicated neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. trichocarpa</em> Salicoid</td>
<td>All genes</td>
<td>(-2.46,-1.34)</td>
<td>(-0.31,-0.18)</td>
<td>(4.08,5.77)</td>
<td>(0.48,0.68)</td>
</tr>
<tr>
<td></td>
<td>≥10 neighbors</td>
<td>(-2.41,-0.84)</td>
<td>(-0.31,-0.12)</td>
<td>(3.21,5.58)</td>
<td>(0.38,0.67)</td>
</tr>
<tr>
<td><em>A. thaliana</em> α</td>
<td>All genes</td>
<td>(-2.40,-1.58)</td>
<td>(-0.15,-0.05)</td>
<td>(4.52,6.97)</td>
<td>(0.54,0.85)</td>
</tr>
<tr>
<td></td>
<td>≥10 neighbors</td>
<td>(-3.31,-0.98)</td>
<td>(-0.26,0.10)</td>
<td>(3.68,9.80)</td>
<td>(0.27,1.21)</td>
</tr>
<tr>
<td><em>A. thaliana</em> β/γ</td>
<td>All genes</td>
<td>(-2.87,-2.14)</td>
<td>(-0.09,0.01)</td>
<td>(5.50,8.42)</td>
<td>(0.55,0.90)</td>
</tr>
<tr>
<td></td>
<td>≥10 neighbors</td>
<td>(-4.18,-2.31)</td>
<td>(-0.31,-0.01)</td>
<td>(7.65,18.14)</td>
<td>(0.88,2.55)</td>
</tr>
</tbody>
</table>
Supplemental Figure 3.1 Kernel density plots of the highest domain odds scores for interacting and non-interacting protein pairs in *A. thaliana* (a), *H. sapiens* (b), and *S. cerevisiae* (c).
Supplemental Figure 3.2 The total number of interactions predicted by ENTS compared to other classifiers. ENTS results are shown as horizontal lines, with colors corresponding to confidence thresholds as follows: 0.55 (yellow), 0.65 (light orange), 0.75 (dark orange), 0.85 (red).
Supplemental Figure 3.3 The total number of genes with predicted interactions from ENTS compared to other classifiers.
Supplemental Figure 3.4 The mean GO biological process and (B) cellular component similarity scores for predicted ENTS interactions. Vertical lines at each confidence level show the total range for randomized networks. (D) Shared GO biological process and (C) cellular component similarity for ENTS as compared to other predicted networks.
Supplemental Figure 3.5 Variable importance for the first 30 most important variables in the *A. thaliana*-trained ENTS classifier. Importance was measured as the mean decrease in OOB classification accuracy during training following permutation of the given variable.
Supplemental Figure 3.6 Variable importance for the first 30 most important variables in the *S. cerevisiae*-trained ENTS classifier.
Supplemental Figure 3.7 Variable importance for the first 30 most important variables in the *H. sapiens*-trained ENTS classifier.
CHAPTER IV

DELETIONS REVEAL THE ONGOING PROCESS OF FRACTIONATION AND ENVIRONMENTAL ADAPTATION IN THE *POPULUS TRICHOCARPA* GENOME
Abstract

Insertion-deletion (indel) polymorphisms are an important, yet understudied, component of the total genetic variation in natural populations. Two-thirds of the total possible indel events within coding sequence (CDS) cause putative nonfunctionalization events through the introduction of frameshifts, while non-CDS mutations may alter or eliminate regulatory sequences. Furthermore, indels may encourage the formation of SNPs during recombination events. In this study, we used whole genome resequencing of 545 individuals to characterize small (< 50bp) and large (>100bp) deletions with respect to the reference genome of the forest tree, *Populus trichocarpa*. We demonstrate that large deletions are primarily enriched in areas of high repetitive content, while the opposite is true of small deletions, which are enriched in areas of high recombination and positively associated with the frequency of SNPs. We find distinct patterns of enrichment between putatively nonfunctionalizing mutations and small non-frameshift mutations, with the former highly enriched among catalytic whole genome duplicates but highly depleted among transcription factors and the latter highly enriched among transcription factors but concentrated outside of DNA-binding regions. Genes affected by nonfunctionalizing mutations also have globally reduced mean expression but increased specificity for the reproductive tissues. Finally, we use indels to detect several regions conferring potential environmental adaptation, many of which could not be detected using only SNPs. These patterns closely parallel the historical patterns of retention and loss following the last whole genome duplication, thereby implicating indel mutations as a mechanism of ongoing fractionation with the paleopolyploid genome.
**Introduction**

Insertions and deletions (indels) comprise an important component of the genetic variation within natural populations. Single mutations may easily abrogate the function of regulatory sequence through obliteration of binding sites or of coding sequences through the introduction of frameshifts or nonsense codons. Yet, studies of indel genetic diversity have largely lagged behind those for single nucleotide polymorphisms (SNPs). Several large-scale surveys of indel variation in humans have been performed in recent years, demonstrating the high frequency of these variants, including hundreds of small indels within the coding regions of healthy individuals (Bhangale et al., MacArthur and Tyler-Smith, Montgomery et al., Mullaney et al.,). Such studies have shown that diversity is reduced in functional genic and non-genic functional regions relative to SNPs (Montgomery et al.,), suggesting a large contribution of these polymorphisms to the overall genetic load. Moreover, human indels are enriched among expression quantitative trait loci (eQTL) of known disease-causing loci (Montgomery et al.,), illustrating their potential impact on common phenotypic variation.

Indel variation may also contribute substantially to functionally relevant variation in plants; indeed, the presence of heterozygous indels is suggested to have a causal relationship with the observed SNP variation (Hollister et al.,). However, comprehensive surveys were infeasible prior to the availability of cost effective high-throughput sequencing technologies. Nonetheless, several studies suggest that indels may explain a high proportion of the variance in agronomically-important quantitative traits and contribute to heterosis, the phenomenon whereby hybrid offspring exhibit superior performance to either parent. In rice, indel polymorphisms are significantly enriched within the promoter regions of genes with additive and transgressive patterns of expression in hybrid crosses (Zhang et al.,). Additionally, studies of Arabidopsis
*thaliana* and *Glycine max* have demonstrated that indels are enriched within clusters of resistance and stress-response genes (Lu et al., McHale et al.,), indicating that they may have important roles in local adaptation to pathogens or environmental stressors.

The propensity of indels to induce pseudogenization of genes within their vicinity makes them particularly interesting for studies of evolution following whole genome duplication, a common phenomenon among flowering plants (Bowers et al., Jiao et al.,). During this time the neopolyploid genome is reduced to a diploid genome with large stretches of intergenomic synteny through a process of large-scale structural rearrangements and smaller-scale resolution of duplicate gene functions via pseudogenization, subfunctionalization, and neofunctionalization (Lockton and Gaut, Mandakova et al., Thomas et al.,). This process, termed fractionation, exhibits a preference for the retention of duplicate genes with putatively higher connectedness within biological networks (Buggs et al., Makino and McLysaght, Rodgers-Melnick et al., Thomas et al.,). The gene balance hypothesis attributes such biases to the potentially deleterious effects of dosage imbalances when the stoichiometry of interacting genes is disrupted, as would be the case if a single, highly-connected whole genome duplicate were pseudogenized or duplicated in isolation from its neighbors (Birchler and Veitia, Freeling and Thomas,). However, the extent to which fractionation is an ongoing process within paleopolyploids and whether the observed biases among retained duplicated genes apply to nonfunctionalizing polymorphisms within populations is currently unclear.

In this study, we discovered a whole genome set of large (>100bp) and small (<50bp) deletion polymorphisms across the natural range of the forest tree *Populus trichocarpa* (Torr. & Gray). We examine the functional content of sequence affected by indel polymorphism, determine the relationship between genic indel content and duplication status, assess the role of network
connectivity on the frequency of segregating indels, and determine whether significant associations exist between indel variants and the environmental variation across the range of *P. trichocarpa.*
Methods

Resequencing data

We obtained plant materials from 1100 black cottonwood (*Populus trichocarpa* Torr & Gray) from core of the species’ range, the Willamette Valley in Oregon, and central British Columbia, as previously described (Slavov et al., ). We resequenced a set of 673 genotypes to 15x depth using the Illumina Genome Analyzer. Short reads were then aligned to the *P. trichocarpa* version 3 genome using BWA 0.5.9-r16 with default parameters (Li and Durbin, ). We ensured optimal mate-pair placement and marked duplicate molecules using the FixMateInformation and MarkDuplicates methods in the Picard package ([http://picard.sourceforge.net](http://picard.sourceforge.net)). Next, we called SNPs and small indels for the merged dataset using SAMtools mpileup (-E –C 50 –DS –m 2 –F 0.000911 –d 50000) and bcftools (-bcgv –p 0.999089) (Li et al., ).

Deletion discovery and genotyping

We performed large deletion discovery and initial genotyping on a set of 88 trees from the latitudinal range of *P. trichocarpa* using the program GenomeSTRiP (Handsaker et al., ) with default settings. We restricted the discovery phase to 88 trees as a trade-off between the computational intensity of this step and the desire to assay most of the common large deletions. We then phased the large indels, along with a panel of SNPs from mpileup with quality scores above 30, using the likelihood imputation method in BEAGLE to obtain posterior probabilities for large deletion genotypes in the 88 trees (Browning and Yu, ). Overlapping deletions of similar size (90% of smaller deletion overlapped by larger) and concordant genotypes (Pearson $r^2 > 0.95$) were merged, resulting in 7,089 deletions with minimal redundancy. This formed our
large deletion discovery set. Finally, we performed genotyping for the deletions in the large deletion discovery set on all 547 trees using the GenomeSTRiP genotyping module.

We validated a randomly chosen set of 16 large deletions using nested PCR assays on a subset of the 88 trees used for deletion discovery. The assays used 1 forward and two reverse primers for each locus, wherein the forward and first reverse primers were located in the flanking regions of the deletion, and a second reverse primer was located within the putatively deleted regions (Supplemental Figure 1).

Following genotyping of the large indels, we combined them with SNPs having quality scores above 30 and small indels with Pearson $r^2$ above 0.95. The Pearson $r^2$ value was based on the correlation between the expected dosage of the alternative allele based on the normalized genotype likelihoods and the values based on assuming the maximum likelihood dosage. We removed any polymorphisms with HWE exact test p-values less than $1 \times 10^{-6}$, and ran likelihood-based phasing and imputation using BEAGLE on the entire set of 547 individuals. Following phasing, we used the smartpca program to identify genetic clusters (Price et al., ). Using only small and large indels, we found similar clusters to those identified using only SNPs (Supplemental Figures 2-3). Using the software GCTA, we identified and removed near relatives with an inbreeding coefficient of greater than 0.075, corresponding to the relatedness of first cousins (Yang et al., ). We also removed two hybrids identified as outliers within the PCA analysis, leaving a total of 545 individuals. We then retained all polymorphisms with BEAGLE $r^2$ of at least 0.95 and a minor allele count of at least 2 for subsequent analyses. This included 10,258,520 SNPs, 254,464 small deletions, and 5,660 large deletions. We excluded the set of small insertions discovered using mpileup due to an excess of homozygosity relative to small
deletions and SNPs that we suspected to be caused by lack of sensitivity for heterozygotes for insertions relative to the reference genome (Supplemental Figure 4).

Analysis of Deletions Affecting Coding Sequences

We further analyzed a subset of the deletion polymorphisms overlapping CDS regions based on the *P. trichocarpa* v. 3 annotation ([www.phytozome.org](http://www.phytozome.org)). We also identified conserved noncoding sequence (CNS) in the *P. trichocarpa* genome relative to *V. vinifera*, *A. thaliana*, *M. trunculata*, and duplicate regions within *P. trichocarpa* using the CNS discovery pipeline (Woodhouse et al., ). CpG islands were defined using the software newcpgseek in the EMBOSS package with default settings. We determined the expected sizes of the small CDS-associated deletions by sampling 10,000 times, with replacement, an equal number of loci from among all small deletions. We determined Z values for the observed counts using the mean and standard deviation of the expected distribution, and calculated 2-sided p-values as the fraction of Z-values in the expected distribution with an absolute value greater than or equal to |Z_{obs}|.

We characterized the median minor allele frequencies (MAF) for different sizes of CDS-associated deletions, grouped by whether they affected gene pairs resulting from the Salicoid whole genome duplication. Within each group, we calculated the median MAF and compared it to the expected range of MAFs for the same number of deletions randomly sampled from the entire set. For the size comparison, we calculated the differences between the observed median for each size and the range of expected values. Similarly, we calculated expected median MAF ranges for deletions grouped by size category (small or large) and duplicate status by resampling among the size category. In each case we resampled from the given category of deletion (small or large) 10,000 times with replacement. Two-tailed p-values were then calculated based on Z-
values calculated using the mean and standard deviation of the expected median MAF
distribution.

Analysis of expression

We obtained a microarray dataset, GSE21484, for *P. trichocarpa* containing expression
data for 16 tissue types (Rodgers-Melnick et al., ). We evaluated the mean over all tissues for
genes affected by small non-frameshift, small frameshift, and large deletions, grouped by
Salicoid duplicate status. We then compared these values to the expected means based on 10,000
resamplings of all genes with the given duplicate status for which expression data was available.
We also performed an ANOVA on the expression values of all genes affected by at least 1
deletion in their CDS regions, using Salicoid duplicate status and the deletion type (large, small
frameshift, or small non-frameshift) as explanatory variables. Due to non-normality among the
residuals, we obtained confidence intervals for all coefficients using a nonparametric bootstrap
with 10,000 iterations.

In addition to comparisons of absolute mean expression, we assessed the relationship
between Salicoid coexpression and deletion. We did this by comparing the mean coexpression of
Salicoid duplicate pairs affected by deletions in at least one of the paralogs to the expected range
of mean co-expression among 10,000 resampled sets of Salicoid duplicates. Coexpression was
calculated using the Pearson correlation coefficient between the expression profiles of the
Salicoid duplicates. We also compared the tissue specificities of genes within a duplicate pair
wherein one member was affected by a given class of deletion (small frameshift, small non-
frameshift, or large) and the other was not. Tissue specificity for each tissue, x, was calculated as
follows:
\[ T S_x = \frac{E_x}{\frac{1}{n} \sum_{y \neq x} E_y} \]

where \( E_x \) is the expression level in tissue \( x \), and \( n \) is the total number of tissues. In each tissue, we then calculated the difference between the tissue specificities of the duplicate by subtracting the value of the unaffected paralog from that of the affected paralog. Following this, we assessed the significance in each tissue by comparing the median observed difference in each tissue with the expected range, where the expected values were calculated by performing the same calculation 10,000 times but with random permutations of which paralog was considered to be affected by the deletion.

**Functional enrichment**

We evaluated differences in gene ontology (GO) functional categories among genes with different types (small frameshift, small non-frameshift, and large) of deletions using Fisher’s exact test. GO categories were taken from the best hit in *A. thaliana* and converted into GOSLIM plant categories using map2slim ([http://search.cpan.org/~cmungall/go-perl/scripts/map2slim](http://search.cpan.org/~cmungall/go-perl/scripts/map2slim)). We performed each enrichment test with the total population either consisting entirely of genes with or without retained Salicoid duplicates. The sample within the population then consisted of all genes with that status that were affected by CDS deletions of the given type. In order to correct for multiple testing while accounting for the hierarchical nature of the GOSLIM ontology, we ran 10,000 permutation tests wherein the enrichment analysis was performed on a randomly selected group of genes having the duplicate status of interest. We then considered categories to be significantly enriched if the observed two-tailed p-value was below the lowest p-values of at least 95% of the permutation tests.
We retained all categories that were considered significant in at least 1 enrichment analysis. We then converted the uncorrected two-tailed p-values to signed Z-values as follows:

\[
Z = \text{sign} \left( \ln \left( \frac{f_s}{f_p} \right) \right) \Phi^{-1} \left( 1 - \frac{p}{2} \right)
\]

where \( \Phi \) is the cumulative distribution function of a standard normal distribution, \( f_s \) is the proportion of genes with deletions annotated with the GO category, and \( f_p \) is the proportion of genes in the population annotated with the GO category. We then performed two-way hierarchical clustering of duplication/deletion status and significant GO categories using single linkage assignment and a Euclidean distance metric based on the vectors of signed Z-values.

**Network-based Analyses**

We used the predicted *P. trichocarpa* protein-protein interaction (PPI) network to test whether genes affected by frameshift mutations have lower connectivity than those unaffected by such mutations. The predicted PPI network was created using the ENTS random forest classifier (Rodgers-Melnick et al., submitted), using an interaction confidence threshold of 0.65. We also tested whether connectivity within the poplarCyc version 3 metabolic network (Zhang et al., ) was affected by frameshift deletions within Salicoid duplicates. We evaluated the connectivity within the metabolic network by creating a graph in which two genes were joined by an edge if one or more products of a reaction catalyzed by the first gene served as reactants for the reaction catalyzed by the second gene. In order to avoid non-meaningful connections based on common compounds such as water and ATP, we excluded compounds present in more than 15 reactions during network construction. For both networks, we assessed significance using a one-tailed Mann-Whitney U test.
**Clustering and depletion analyses**

We further investigated a group of transcription factors affected by an exceptional number of small non-frameshift deletions. We first identified groups of homologous transcription factors by running a BLASTP of all proteins annotated with GO:0045449, regulation of transcription, against all proteins in the *P. trichocarpa* genome with an E-value threshold of $1 \times 10^{-5}$. We created an undirected graph of edges among proteins with hits to one another and then aligned the proteins affected by small non-frameshift deletions from the largest groups containing domains with annotated DNA-binding activity. This included families of myb-like, HLH, and WRKY proteins. We performed protein alignments using MUSCLE with default parameters (Edgar, ).

Following alignment, we tested for depletion of small non-frameshift deletions in the DNA binding regions of each family by randomly permuting the starting locations of the deletions over the positions in each protein and subsequently counting the number of deletions in the permutation that overlapped the DNA-binding domain(s). We performed this permutation 10,000 times for each group and calculated a p-value as the fraction of permutations in which the number of deletions overlapping the DNA binding region was less than or equal to the observed value. We also tested for significant clustering using a permutation test by calculating whether the sum of distances between deletions in the alignment that were nearest neighbors to one another was significantly smaller than expected by chance.

**Association Analyses**

We obtained climate data from the WORLDCLIM 2.5-minute GIS layer using DIVA-GIS version 7.5 (Hijmans et al., ). We then narrowed the set of individuals to the 489 not removed as
outliers during smartpca with default settings (6.0 standard deviations along top 10 components), which eliminated the trees from California and the Willamette valley. We also narrowed the sets of SNP and indel markers to those with MAFs of greater than 0.025, leaving 165,203 deletions and 5,402,341 SNPs. Using all environmental variables in the WORLDCLIM dataset at the coordinates of the 489 remaining trees, we formed principal components (PCs) following the normalization of each variable. The first 4 climate PCs were then used as responses for association analysis with EMMAX (Kang et al., ). We ran two sets of association analysis, the first using only indel genotypes and the second using only SNP genotypes. In each case, we included as covariates all PCs from the genotype data with Pearson correlation p-values of less than 0.005 with the climate PC response variables. We also accounted for small-scale structure using a Balding-Nichols kinship matrix estimated by EMMAX based on the 5.4 million SNPs or 165,203 deletions for SNPs and deletions, respectively (Kang et al., ). We accounted for multiple testing for indel associations by performing 10,000 repetitions of each analysis, wherein we randomly permuted response values at each iteration. Following each permutation, we retained the lowest p-value, using the lowest 5% of p-values to set the significance of each analysis.
Results

Characterization of small and large deletions

We discovered a total of 5,660 large deletions and 254,464 small deletions that were present in at least 2 copies within a population of 545 *P. trichocarpa* trees from across the latitudinal range of the species. Due to the difficulty of identifying large deletions using high throughput sequencing data, we also validated genotypes for 16 randomly chosen loci using a PCR assay (Supplemental Table 1). All assayed loci carried the band sizes expected given the predicted deletion sizes, and match rates between the assayed and predicted genotypes ranged from 82% to 100%, with over half of the loci matching at rates above 90%. Furthermore, the large majority of genotyping mismatches occurred when the PCR assay failed to detect the allele that would produce the larger band in predicted heterozygotes. Therefore, PCR scoring error most likely accounts for most mismatches.

Large deletions cover a total of 26.4 Mb and range in size from 205 bp to 470.3 kb, with a median of 1.3 kb. Small deletions cover a total of 1.24 Mb and range in size from 1 to 47 bp, with a median of 1 bp. Both large and small deletions have MAFs ranging from 0.00183 to 0.5, with medians of 0.058 and 0.199, respectively. The content of the deleted sequences also varies substantially between the two sets of deletions. Most strikingly, the large deletions are highly enriched for repeat elements, which cover over 70% of the large deletion space, nearly twice as much as the 38% of the total genome covered by repeats (Figure 1A). This enrichment is present for nearly every class of transposable element (TE) identified by RepeatMasker, which together cover approximately half of the large deletion space. Indeed, over 30% of the total sequence affected by large deletion events is annotated for the presence of the gypsy family of LTR retrotransposons. By contrast, TEs are underrepresented among small deletions, constituting
approximately 10% of the total small deletion space compared to the 20% of the genome containing TE sequences. Strikingly, both small and large deletions have an underrepresentation of rRNA sequence.

Large and small deletions also differ with respect to putatively functional elements, including both genic and conserved noncoding sequence (CNS) (Figure 1B). We find that small deletions tend to be enriched in and near functional elements: a linear model of small deletion content in 250kb windows with gene content, SNP content, and CpG island content as predictors explains over 82% of the variance between windows, with all explanatory variables having positive effects (Table 1; Supplemental Table 2). On a more local scale, we observe a strong enrichment for small deletions surrounding recombination hotpots (p < 0.0001; permutation test) (Supplemental Figure 5). Intriguingly, we also find an enrichment of SNPs within the first 150bp surrounding small deletions (Figure 1D), though SNPs are depleted among the first 10bp due to the downgrading of BAQ scores by mpileup in regions with nearby indels (SAMtools, ). Moreover, this enrichment is enhanced for common deletions (MAF > 0.25) compared with rare deletions (MAF < 0.005).

In contrast to small deletions, large deletions have an underrepresentation of all functional classes. Among genes, this underrepresentation is more extreme for retained duplicates of the Salicoid WGD than for genes without retained Salicoid duplicates (hereafter referred to as non-Salicoids) (Figure 1C). Moreover, the paucity of CNS in sequences affected by large deletions is more extreme for elements shared across multiple species compared to those that are only conserved with respect to a single specie

*CDS deletions have biased size and minor allele frequency distributions*
We characterized the size distribution of small deletions contained within coding sequences. Although 1bp deletions still comprise the mode of the distribution, these frameshift deletions are significantly underrepresented relative to the number expected given the background size distribution of small deletions (Figure 2A). Conversely, CDS small deletion sizes that correspond to multiples of three, and thus would not cause frameshifts, are significantly enriched relative to the numbers expected by chance.

The minor allele frequencies of small deletions also show a pattern emphasizing multiples of three. While the median minor allele frequency (MAF) is lower than the background expectation across both large and small deletions encompassing CDS, the differences between the observed and expected median MAFs are narrower for non-frameshift small deletions than frameshift deletions of similar size (Figure 2B-C). Nonetheless, these non-frameshift mutations do still tend to be at lower MAF than expected.

*Genes with CDS deletions are less expressed*

All genes with small frameshift and large deletions abrogating their CDS regions have significantly lower than expected mean expression (Supplemental Figure 6A). Moreover, there is a significant negative correlation between mean expression levels and MAFs for non-Salicoids ($\rho=-0.0686$, $p=0.0052$), meaning that, in general, the frequency of the deletion polymorphism is negatively correlated with expression level, as expected. However, the same is not true for genes with retained duplicates ($\rho=-0.0049$, $p=0.855$). Retained Salicoid duplicates with small non-frameshift deletions also have significantly lower mean expression. However, non-Salicoid genes affected by small non-frameshift deletions have much less of a difference between observed and expected expression. The reduced expression extends across all tissues tested for each category
with a significant difference (Supplementary Figures 6-11). We also find that, controlling for
duplicate status, genes with non-frameshift deletions have significantly higher expression than
those with frameshift deletions, while genes with large deletions have significantly lower
expression than genes with small frameshift deletions (Table 2).

Salicoid duplicate pairs also have lower co-expression than expected when at least one of
the genes was affected by a small frameshift deletion (Supplemental Figure 6B). Additionally,
duplicates affected by frameshift mutations have significantly more concentrated expression in
the catkins and post-pollination female floral tissues and significantly less concentrated
expression in the axillary buds, shoot apex, and roots compared to the duplicates not affected by
small frameshifts (Figure 3A). Although we do not find evidence of reduced co-expression
between Salicoid duplicate pairs affected by large deletions, we do find that duplicates affected
by large deletions have expression significantly more concentrated within the mature catkin and
female floral bud initials but less concentrated within the mature leaves and phloem (Figure 3B).
In contrast to the frameshifts and large deletions, we find no significant reductions in co-
expression or tissue specificities between Salicoid duplicate pairs affected by small non-
frameshift deletions.

*Non-frameshift deletions have distinct functional enrichment*

Hierarchical clustering of enrichment of functional categories resulted in a clear
separation of small non-frameshift deletions from small frameshifts and large deletions (Figure
4). Most strikingly, genes involved in nucleic acid binding, which include a large number of
transcription factors, are highly enriched for the presence of non-frameshifting deletions. By
contrast, these deletions are underrepresented among transporters and catalytic genes,
particularly among those with retained Salicoid duplicates. Furthermore, these same categories that are most enriched among genes affected by small non-frameshift deletions are underrepresented for frameshift and large deletions. Likewise, genes with catalytic activity tend to be enriched for large and small-frameshift deletions, particularly among retained Salicoid duplicates, while having an underrepresentation among all small non-frameshift deletions. This contrasting pattern between the two groups is not universal, as both ribosomal proteins and structural molecules have an underrepresentation of all deletions except for large deletions among retained Salicoid duplicates.

Intriguingly, while genes with and without Salicoid duplicates generally have similar patterns of non-frameshift enrichment, they have contrasting patterns for several categories among frameshift deletions. For instance, although plastid-associated genes are significantly underrepresented among non-Salicoid genes with small frameshift deletions, they are overrepresented among frameshifted Salicoid genes. However, the significance of the Salicoid enrichment does not hold under a correction for multiple testing. Likewise, while Salicoid duplicates with catalytic activity are highly enriched for small frameshift deletions, non-Salicoids are no more enriched than expected by chance.

*Catalytic Salicoid duplicates have fewer metabolic connections*

We find that both Salicoid and non-Salicoid genes affected by small frameshift deletions have significantly lower connectivity within the predicted *P. trichocarpa* PPI network than genes not affected by such putatively nonfunctionalizing deletions (Salicoid p = 0.0215; Non-Salicoid p = 0.00450; One-tailed Mann-Whitney U test), suggesting that network centrality affects the extent of purifying selection. Due to the high overrepresentation of genes with catalytic activity
among retained Salicoid duplicates, we hypothesized that these may be more prone to elimination under an ongoing fractionation process within the *P. trichocarpa* genome. Under the predictions of the gene balance hypothesis, we would expect the genes with frameshifts to trend toward less dosage sensitivity than non-frameshifted catalytic Salicoid duplicates, potentially due to less central roles within the *P. trichocarpa* metabolic network. We used the poplarCyc metabolic network to assess this hypothesis, by testing whether the catalytic Salicoid genes with frameshifts share a significantly lower number of compounds with other genes in the network than the catalytic Salicoid genes without frameshift deletions. Indeed the median degree for genes with frameshift deletions is 37, while that for genes without frameshift deletions is 50, a significant difference (*p* = 0.0357; one-sided Mann-Whitney U) (Figure 5A).

Non-frameshift deletions in nucleic acid binding genes are depleted in DNA binding regions

Our results demonstrate that while transcription factors and other nucleic acid binding genes are underrepresented among genes with frameshift deletions, they may have a strong propensity for diversification through non-frameshift deletions. Interestingly, there is a significant enrichment of nucleic acid binding genes for which both Salicoid duplicates have small non-frameshift deletions (*p* < 0.01) (Figure 5B). Furthermore, although there is a general enrichment of CNS in intergenic regions adjacent to nucleic acid binding genes, there is a significant underrepresentation of CNS conserved with respect to grape flanking Salicoid nucleic acid binding genes affected by frameshift deletions (*p* = 0.001; 2-tailed Fisher’s exact test). Two other classes of CNS – those conserved with respect to *M. trunculata* and *P. trichocarpa* and those conserved with respect to those two species in addition to *V. vinifera* - also had suggestive underrepresentations of CNS flanking the frameshifted Salicoid nucleic acid binding genes (FDR = 0.1).
The enrichment of non-frameshift deletions in nucleic acid binding genes prompted us to question whether particular regions of these proteins were disproportionately affected, perhaps due to varying degrees of purifying or positive selection along the lengths of these proteins. We tested for significant spatial clustering of non-frameshift deletions and whether such deletions are significantly depleted from the highly conserved DNA binding domains that define the myb-like, bHLH, and WRKY groups of transcription factors. Although we find no evidence of significant spatial clustering in any group, we do find significant depletions of small deletions in the DNA binding regions of myb-like TFs ($p < 1 \times 10^{-4}$) and WRKY TFs ($p = 6 \times 10^{-4}$) (Figure 5C; Supplemental Figures 13-14). Furthermore, a bHLH TF group that does not achieve statistical significance in our analysis ($p = 0.11$) has deletions restricted to the periphery of the DNA binding domain.

Several deletions are highly associated with environmental variation

In order to examine the relationships between genotypes and environmental factors, many of which have strong correlations with one another, we divided the environmental variation into principal components (Supplemental Figures 15-16; Supplemental Table 3). After testing the first 4 environmental PCs for associations with both deletions and SNPs, we found a single deletion having a significant association with the first PC and 4 deletions significantly associated with the fourth PC (Figure 6; Supplemental Figures 18-20). As a liberal basis of comparison, we applied the same multiple testing criterion to SNPs, resulting in 8 significant SNP associations for the first PC and 42 for the fourth. Altogether, the deletion-only analysis implicates 10 genes in environmental associations over all 4 PCS, compared to the 74 implicated by the SNP-only analysis (Supplemental Figure 17), where we considered a gene to be implicated if a significant
marker fell within the genic region or in the region flanking it. Surprisingly, only two genes were implicated by both analyses.

The first environmental PC, which contrasts several measures of temperature with the precipitation during the warmest and driest months, is associated with a deletion residing in an intron of Potri.001g07330. The *A. thaliana* homolog of this gene, *DIS2* (AT1G30825), is an Arp2/3 complex subunit protein involved in the trichome maturation process. Although this region is not significantly associated with the first PC under the SNP-based analysis, the second most significant region for the SNP analysis occurs within a HECT ubiquitin-protein ligase with an *A. thaliana* ortholog, *UPL4* (AT5G02880), also involved in trichome development (Downes et al., ).

The fourth environmental PC primarily contrasts seasonal variation in precipitation with the mount of precipitation received during the warmest and driest months. The first two associated deletions with this PC are overlapping large deletions within an intergenic region flanking Potri.018g044100, a carotenoid cleavage dioxygenase and homolog of *A. thaliana MAX4* (AT4G32810). This region is also strong implicated by 28 of the 42 significant SNPs. The remaining two significantly associated deletions for PC4 consist of small deletions residing on separate chromosomes. The first is an intergenic deletion proximal to the 3’ end of Potri.008g086000, a HOX domain protein that is predominantly expressed in the male and female floral initials of *P. trichocarpa*. The second is a frameshift deletion within Potri.001g272900, a putative proton-dependent oligopeptide (POT) transporter that is also implicated by the SNP-only analysis.
Table 4.1. Linear model explaining small deletion frequency in 250kb windows across the genome

|                            | Estimate  | Std. Error | t-value | P(>|t|)  |
|---------------------------|-----------|------------|---------|----------|
| SNP Content               | 1.31×10^{-3} | 1.75×10^{-5}  | 75.11   | <2×10^{-16} |
| Gene Content              | 0.14      | 3.5×10^{-4}  | 40.06   | <2×10^{-16} |
| CpG Content               | 1.6×10^{-3}  | 5.73×10^{-4} | 2.81    | 0.0049   |

Residual Standard Error: 1.45 on 1565 degrees of freedom
Multiple R^2: 0.827
F-statistic: 2495 on 3 and 1565 degrees of freedom, p-value: < 2.2×10^{-16}
Table 4.2 95% confidence intervals for the effects of Salicoid duplicate status and deletion type on expression level. Duplicate status effect is relative to non-Salicoid duplicates, while deletion type effects are relative to frameshift deletions. Confidence intervals were generated by 10,000 bootstrapped samples.

<table>
<thead>
<tr>
<th></th>
<th>95% CI Lower</th>
<th>95% CI Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salicoid Duplicate</td>
<td>0.0857</td>
<td>0.2533</td>
</tr>
<tr>
<td>Large Deletion</td>
<td>-0.5329</td>
<td>-0.2318</td>
</tr>
<tr>
<td>Non-frameshift Deletion</td>
<td>0.1004</td>
<td>0.2714</td>
</tr>
</tbody>
</table>
Figure 4.1 Cumulative amount of sequence space within the large deletions, small deletions, and genome composed of (A) repeat elements, (B) genes and conserved noncoding sequence, and (C) genes with and without retained Salicoid duplicates. (D) SNP density within the region adjacent to small deletions.
Figure 4.2 Size and MAF distributions of small deletions within CDS regions. (A) Observed (black) and expected (grey) counts of CDS small deletions. (B) The range of differences between the observed median MAF and the range of expected MAFs given the background distribution for that deletion size. (C) Observed and expected MAFs for small and large deletions overlapping genes with and without retained Salicoid duplicates. Error bars indicate 95% confidence intervals. (*** p < 0.001, ** p < 0.01, * p < 0.05; Two-tailed)
Figure 4.3 Differences in median tissue specificity between Salicoid duplicates in which one member of the pair is affected by (A) a small frameshift deletion or (B) a large deletion. Blue dots indicate the observed median difference between the tissue specificity values, and grey bars indicate 95% confidence intervals for no true difference in specificity.
**Figure 4.4** Enrichment of plant GO-SLIM categories among genes with deletions affecting their CDS. Highly positive Z-values (red) indicate an enrichment of the category, while highly negative Z-values (blue) indicate an underrepresentation of the category.
Figure 4.5 (A) Bootstrapped median degrees for Salicoid catalytic genes in the poplarCyc metabolic network. (B) Observed (black) and expected (grey) numbers of Salicoid duplicate pairs among nucleic acid binding genes containing small non-frameshift deletions. (C) Positions of small non-frameshift deletions within a group of myb-like transcription factors. Horizontal lines indicate sequences within a multiple sequence alignment, while red triangles indicate the positions of the deletions. Sequence conservation at each position is given by the vertical lines at the top, and the DNA-binding myb repeats are highlighted in blue.
Figure 4.6 (A) QQ-plot and (B) Manhattan plot of deletion associations with the 4\textsuperscript{th} environmental PC. Dotted horizontal lines indicate p = 0.05, as assessed by 10,000 permutations.
Discussion

*Biases in deletion content reflect mechanisms of formation*

We find that large deletions are enriched in genomic regions with high TE content but depleted from regions containing both genes and CNS. This result is consistent with previous data from *A. thaliana* (Ziolkowski et al., ). It supports a model whereby large deletions may be favored in high TE regions through unequal recombination between neighboring elements or intrachromosomal recombination between flanking repeats of single elements. Moreover, this model also explains the relative dearth of both genes and intergenic CNS within regions spanned by large deletions, as TEs accumulate in gene-poor areas due to purifying selection for the maintenance of gene expression (Wright et al., ).

Surprisingly, we observe an enrichment of both genic and CNS sequence among regions overlapped by small deletions. Although this enrichment is not as extreme as the relative depletion of large deletions in putative functional regions, it does suggest that purifying selection does not promote a reduction in the frequency of all deletion types. Moreover, the positive predictors of small deletion content in our analysis – SNP content, gene content, and CpG island content – were all strong positive predictors in an earlier study of recombination in *P. trichcarpa* (Slavov et al., ), suggesting that increased recombination within regions of high gene content may also promote the formation of small deletions. This supposition is consistent with our results, showing an enrichment of small deletions within and surrounding recombination hotspots. It is also consistent with an analysis in vertebrates demonstrating both increased gene content and short deletion rates within areas of high recombination (Nam and Ellegren, ).

Furthermore, the connection between deletions and recombination is also mechanistically favored by the lower fidelity and strong deletion bias of DNA polymerase δ (Fortune et al., ), the
enzyme most often utilized during the extension of the heteroduplex formed during eukaryotic recombination (Maloisel et al., ). Despite this, recent results in humans demonstrated a lack of any enrichment for indels within areas surrounding recombination hotspots and implicated polymerase slippage as the major source of indel genesis (Montgomery et al., ). However, vertebrates and plants are known to have contrasting patterns of microsatellite evolution, with plants having a negative association between microsatellite content and genome size (Ellegren, ). Therefore, our results may reflect a fundamental distinction between the major mechanisms of small indel generation in plants and animals. However, we also note that the placement of hotspots within regions of high gene content and low-complexity sequence makes disentangling the effects of genomic context and recombination difficult.

Recombination in areas surrounding small indels when in the heterozygous state is also implicated in the increased generation of SNP polymorphisms (Hollister et al., , Tian et al., ), and this effect should be enhanced among obligate outcrossers such as P. trichocarpa (Hollister et al., ). In support of this mechanism, we find that regions surrounding small deletions have elevated SNP density, and that this effect is greater for deletion loci with high MAF. Altogether, we believe our analysis supports a model whereby small deletions are both promoted by recombination and, in turn, further promote higher SNP density.

*The size and MAF distributions of deletions reveal the spectrum of purifying selection*

Deletions overlapping CDS regions have high potential for causing severe biochemical consequences, as 2/3 of all possible deletions will introduce frameshifts, likely yielding nonfunctional protein products or, more likely, degradation of transcribed mRNAs through nonsense mediated decay (Frischmeyer and Dietz, ). Our results highlight the relative influence
of purifying selection on frameshift versus non-frameshift deletions, as small non-frameshift deletions are more common in the genome than expected given the background size distribution of small deletions. Moreover, the reduced difference between the observed and expected MAFs of small non-frameshift deletions suggests less stringent purifying selection. However, even non-frameshift deletions tend toward lower than expected MAF, similar to the more extreme patterns of both small frameshift and large deletions. We therefore expect that CDS deletions contribute substantially to the genetic load.

*Lower expression and less connectivity increase the probability of deletion polymorphisms*

Our analysis of *P. trichocarpa* microarray data indicates that genes harboring deletions also tend to have significantly reduced expression across all tissues and that this effect is most extreme for large deletions and least extreme for small non-frameshift deletions. We do not believe this relationship is due to deletions directly reducing gene expression, as most of the array data is taken from the tree used to generate the reference genome, Nisqually-1 (Rodgers-Melnick et al., ). Therefore, since the analysis is restricted to loci that are deletions relative to the reference genome, at least one full-length allele should be present for each gene. Altogether, we believe these results suggest that genes with lower expression levels experience diminished purifying selection, a conclusion supported by research in yeast and mammals (Drummond et al., , Hudson and Conant, , Pal et al., ).

The higher expression of genes impacted by small non-frameshift deletions is consistent with the dependence of the CDS small deletion MAF distribution on deletion size, as the expression levels of these genes tend to be lower than expected given their duplication status but still higher than the genes affected by the putatively more severe frameshift and large deletions.
This suggests that small non-frameshift deletions are, on average, slightly deleterious. While we expected genes impacted by small frameshift and large deletions to have lower than average expression, we were surprised to discover significantly lower expression in genes affected by large deletions compared to those harboring small frameshifts. We suspect this difference is influenced by the biases in the locations of large deletions, as genomic regions prone to large deletions are also enriched in TEs and therefore likely to reside within condensed heterochromatin under most circumstances (Lippman et al., ). Furthermore, these regions may be more likely to contain genes subject to relaxed purifying selection and, correspondingly, highly reduced expression (Wright et al., ).

In addition to the globally reduced expression of genes affected by deletions, Salicoid duplicates affected by frameshift or large deletions show evidence of divergent expression patterns. This is most apparent for small frameshifts, where duplicate pairs have significantly reduced co-expression. However, we also find that duplicates impacted by these putatively nonfunctionalizing mutations tend to have more tissue specificity in the reproductive tissues compared to their non-impacted paralogs, particularly within the male catkins. This result supports the recent finding that rapid evolution of genes with high male reproductive tissue specificity in plants is driven by relaxed purifying selection rather than positive sexual selection (Yang and Wang, ). Thus, the continued fractionation of retained Salicoid duplicates may be driven in part by the preferential loss of duplicates that have their expression predominantly confined to the reproductive tissues, perhaps following divergence from broader expression patterns through subfunctionalization, neofunctionalization, or both.

*Patterns of functional enrichment mirror patterns of duplicate retention*
Our analysis of functional enrichment reveals strong patterns of over- and under-representation among different types of deletions. We find similar patterns of enrichment among small frameshift and large deletions, underscoring the potential of these deletions to cause nonfunctionalization. In particular, we find that transcription factors and other nucleic acid binding proteins are significantly underrepresented among the genes containing these types of deletions. This trend toward high purifying selection for the preservation of TF function parallels the high retention of these genes following whole genome duplications in *P. trichocarpa* and a number of other species (Buggs et al., Rodgers-Melnick et al., Thomas et al., ). Furthermore, the significant underrepresentation of CNS among Salicoid nucleic acid binding genes affected by frameshifts suggests that the loss of genes with conserved regulatory motifs is disfavored. We also suspect that this is most apparent for grape CNS due to the lack of subsequent WGDs and associated fractionation events in the *V. vinifera* lineage following the ancestral eudicot triplication (Jiao et al., ).

In light of the apparent purifying selection for nucleic acid binding gene function, we were surprised to find significant enrichments of small non-frameshift deletions among these same groups of genes. Many of these genes, such as the MYB and bHLH TFs, are members of large plant-specific gene families that combinatorially regulate processes such as epidermal cell differentiation, flavonoid biosynthesis, and phytochrome signaling (Feller et al., ). Our analyses suggest that deletions are largely restricted away from the DNA binding regions, implying that nucleic acid binding proteins are prone to the accumulation of small length changes outside of these conserved regions, either through nearly neutral processes or through positive selection, perhaps due to diversification of protein-protein interactions between co-regulators. Although the diversification of TFs through small deletions appears largely unexplored in the literature, there
is evidence for the importance of indels in the evolution of interaction specificity among plant MADS TFs (van Dijk et al., ). This suggests that further study of non-frameshift deletions within TFs may lead to insight into the evolutionary dynamics of transcriptional regulatory networks.

In contrast to the pattern of enrichment among nucleic acid binding genes, we found a strong enrichment of catalytic genes among Salicoid duplicates affected by small frameshift deletions. This also mirrors the results of our previous analysis of duplicate genes in *P. trichocarpa*, as catalytic activity is associated with a reduced probability of retention following WGD (Rodgers-Melnick et al., ). This suggests that the deletion polymorphisms are signatures of an ongoing fractionation process whereby duplicate genes in some categories are undergoing pseudogenization. Nevertheless, several thousand Salicoid duplicates are still annotated with catalytic activity, and they have functions in myriad biochemical pathways.

Our analyses of network connectivity indicate that, globally, genes with frameshift deletions tend to have less central roles within PPI networks and, in particular, catalytic Salicoid genes with frameshift deletions tend to have less central roles within metabolic networks. This dependency on network connectivity suggests that dosage sensitivity may have a strong role in determining whether purifying selection will eliminate non-functionalizing mutations to Salicoid catalytic genes. Indeed, an analysis of the *A. thaliana* metabolic network provides evidence for both clustered retention of WGD duplicates, indicating the importance of balanced dosages between interactors in the network, and the biased retention of duplicates catalyzing reactions with high metabolic fluxes, signifying the importance of absolute dosage (Bekaert et al., ). The reduced connectivity of nonfunctionalized catalytic Salicoid duplicates in our analysis may stem from reduced purifying selection for maintenance of duplicates with diminished absolute dosage or relative dosage constraints. However, we do find a suggestive reduction in the proportion of
neighbors with retained Salicoid duplicates (frameshift median 0.48, non-frameshift median 0.54, \( p = 0.064 \); one-tailed Mann-Whitney U), which suggests that catalytic Salicoid duplicates with less constraint to maintain stoichiometric balance between duplicated neighbors may experience reduced purifying selection for maintenance of both paralogs.

*Deletions reveal regions of environmental adaptation*

Although linkage disequilibrium within *P. trichocarpa* is higher than previously assumed for outcrossing forest trees (Slavov et al., ), even moderate LD decay may severely weaken the ability to detect significant associations with genotypes if sample sizes are small and causal variants are not directly assayed. Our indel-environmental association analysis discovered several candidates for environmental adaptation, including several that could not be detected using 5.4 million SNPs. For instance, the region containing Potri.001g073300, a gene implicated in the development of trichomes, was found to be associated with the first environmental PC using indels but not using SNPs. However, a separate region containing a *P. trichocarpa* homolog to *A. thaliana UPL4* (AT5G02880) is highly significant under the SNP analysis but not implicated by any indels. Intriguingly, this HECT ubiquitin-protein ligase is also involved in trichome development (Downes et al., ). Therefore, the use of both SNPs and indels implicates two independent regions with roles in the development of trichomes, which are known to confer drought resistance by protecting against damaging UV radiation and leaf dessication (Choinski and Wise, , Skaltsa et al., ). This adaptive role in drought tolerance is further supported by the first PC’s interpretation as a contrast between hot, dry and cool, wet climates.

Our indel results from the 4th PC also implicates a region that was not associated through the SNP-only analysis. However, the region most strongly associated with the PC was highly
significant under both the SNP and indel analyses. Furthermore, the associated region contains one of the \textit{P. trichocarpa} MAX genes, which have been previously associated with branching and shoot phenology phenotypes using QTL analysis (Brunner et al., ). These phenotypes can have a positive impact on the total biomass production of commercial trees, making this region an excellent target for further study.

\textit{Conclusion}

Our results demonstrate that indel polymorphisms comprise an important component of the total genetic variation in \textit{P. trichocarpa}. Both small and large deletions contribute heavily to the genetic load of the species but may also provide the basis for environmental adaptations such as drought tolerance. Furthermore, our study of deletion density throughout the genome lends new insight into the genesis of genetic variation, as our findings support positive associations between the frequency of recombination, the generation of deletions, and the introduction of SNPs. We find that indels are also a potent long-term evolutionary force, providing a mechanism for the continued biased fractionation of paralogs retained from the most recent WGD and a plausible means for the ongoing diversification of transcriptional networks. Together, we believe our results promote an expanded role for the study of indel variation in whole genome surveys of genetic variation.
Acknowledgements

This work was supported 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, and through a United States Department of Agriculture National Institute of Food and Agriculture pre-doctoral fellowship.
References


**Supplemental Table 4.1** Results of PCR validation assays. Column headings titled as XX/XX indicate PCR result/prediction. The insertion and deletion alleles are represented by the letters “I” and “D”, respectively.

<table>
<thead>
<tr>
<th>Locus</th>
<th>% Match</th>
<th>II/II</th>
<th>ID/ID</th>
<th>DD/DD</th>
<th>II/II</th>
<th>II/DD</th>
<th>ID/II</th>
<th>ID/DD</th>
<th>DD/II</th>
<th>DD/ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr05:24492687-24493101</td>
<td>100%</td>
<td>57</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr09:3171420-3171740</td>
<td>99%</td>
<td>15</td>
<td>34</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr09:4872204-4872596</td>
<td>98%</td>
<td>64</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr01:3801761-3802441</td>
<td>97%</td>
<td>23</td>
<td>35</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Chr10:7441991-7448907</td>
<td>96%</td>
<td>2</td>
<td>14</td>
<td>59</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Chr12:14698332-14699612</td>
<td>96%</td>
<td>24</td>
<td>33</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr10:16531092-16531352</td>
<td>94%</td>
<td>28</td>
<td>30</td>
<td>17</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr01:1216977-1217195</td>
<td>93%</td>
<td>4</td>
<td>15</td>
<td>55</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Chr12:132344-132741</td>
<td>93%</td>
<td>29</td>
<td>29</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Chr03:10497024-10497274</td>
<td>89%</td>
<td>7</td>
<td>35</td>
<td>29</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr18:1892999-1893560</td>
<td>87%</td>
<td>12</td>
<td>30</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr05:7012136-7014123</td>
<td>87%</td>
<td>8</td>
<td>24</td>
<td>28</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr10:17077665-17079354</td>
<td>87%</td>
<td>12</td>
<td>23</td>
<td>12</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr10:20727892-20728581</td>
<td>86%</td>
<td>42</td>
<td>20</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chr04:17754901-17755788</td>
<td>84%</td>
<td>31</td>
<td>24</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Chr01:49510513-49510888</td>
<td>82%</td>
<td>48</td>
<td>12</td>
<td>2</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Supplemental Table 4.2** Spearman correlations (below diagonal) and p-values (above diagonal) between variables in 250kb windows across the genome. Small deletions, large deletions, SNPs, gene content, and CpG islands indicate the counts for those elements, while GC content and N content represent the percentage of GC and N nucleotides within the window.

<table>
<thead>
<tr>
<th></th>
<th>Small Deletions</th>
<th>Large Deletions</th>
<th>SNPs</th>
<th>GC Content</th>
<th>Gene Content</th>
<th>CpG Islands</th>
<th>N Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Deletions</td>
<td>-</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9477</td>
<td>0.0000</td>
</tr>
<tr>
<td>Large Deletions</td>
<td>0.12</td>
<td>-</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>SNPs</td>
<td>0.6</td>
<td>0.32</td>
<td>-</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>GC Content</td>
<td>-0.25</td>
<td>-0.18</td>
<td>-0.51</td>
<td>-</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Gene Content</td>
<td>0.39</td>
<td>-0.14</td>
<td>-0.19</td>
<td>0.19</td>
<td>-</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>CpG Islands</td>
<td>0.00</td>
<td>-0.12</td>
<td>-0.34</td>
<td>0.59</td>
<td>0.35</td>
<td>-</td>
<td>0.0213</td>
</tr>
<tr>
<td>N Content</td>
<td>-0.56</td>
<td>-0.19</td>
<td>-0.43</td>
<td>0.11</td>
<td>-0.32</td>
<td>-0.06</td>
<td>-</td>
</tr>
</tbody>
</table>
Supplemental Table 4.3 Pearson correlations of environmental PCs with individual environmental variables at the sampling locations. The most significant positive and negative correlations for each PC are highlighted with bold boxes.

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Temp. of Coldest Quarter</td>
<td>0.9088</td>
<td>-0.2063</td>
<td>0.1558</td>
<td>0.1776</td>
</tr>
<tr>
<td>Annual Mean Temp</td>
<td>0.9015</td>
<td>-0.1571</td>
<td>0.2376</td>
<td>0.03937</td>
</tr>
<tr>
<td>Mean Temp of Wettest Quarter</td>
<td>0.8857</td>
<td>-0.1978</td>
<td>0.1462</td>
<td>0.1938</td>
</tr>
<tr>
<td>Isothermality</td>
<td>0.8647</td>
<td>-0.3633</td>
<td>0.02176</td>
<td>0.1501</td>
</tr>
<tr>
<td>Max Temp of Warmest Month</td>
<td>0.8278</td>
<td>-0.4093</td>
<td>0.3639</td>
<td>-0.1109</td>
</tr>
<tr>
<td>Precip of Driest Month</td>
<td>-0.8219</td>
<td>0.7023</td>
<td>-0.253</td>
<td>-0.4208</td>
</tr>
<tr>
<td>Mean Temp of Warmest Quarter</td>
<td>0.7861</td>
<td>-0.2422</td>
<td>0.3952</td>
<td>-0.159</td>
</tr>
<tr>
<td>Mean Monthly Temp Range</td>
<td>0.785</td>
<td>-0.5147</td>
<td>0.09635</td>
<td>-0.003049</td>
</tr>
<tr>
<td>Min Temp of Coldest Month</td>
<td>0.7778</td>
<td>0.01097</td>
<td>0.1872</td>
<td>0.1471</td>
</tr>
<tr>
<td>Mean Temp of Driest Quarter</td>
<td>0.7606</td>
<td>-0.25</td>
<td>0.3664</td>
<td>-0.2003</td>
</tr>
<tr>
<td>Precip of Driest Quarter</td>
<td>-0.6687</td>
<td>0.8008</td>
<td>-0.2947</td>
<td>-0.457</td>
</tr>
<tr>
<td>Precip of Warmest Quarter</td>
<td>-0.6367</td>
<td>0.8518</td>
<td>-0.2716</td>
<td>-0.3691</td>
</tr>
<tr>
<td>Precip Seasonality</td>
<td>0.4959</td>
<td>-0.2223</td>
<td>0.7968</td>
<td>0.375</td>
</tr>
<tr>
<td>Annual Precip</td>
<td>-0.2397</td>
<td>0.9547</td>
<td>0.2286</td>
<td>-0.1358</td>
</tr>
<tr>
<td>Temp Annual Range</td>
<td>0.1058</td>
<td>-0.5702</td>
<td>0.4517</td>
<td>-0.2682</td>
</tr>
<tr>
<td>Temp Seasonality</td>
<td>-0.104</td>
<td>-0.224</td>
<td>0.5839</td>
<td>-0.2983</td>
</tr>
<tr>
<td>Precip of Wettest Month</td>
<td>-0.03099</td>
<td>0.8119</td>
<td>0.5031</td>
<td>0.002824</td>
</tr>
<tr>
<td>Precip of Coldest Quarter</td>
<td>0.02657</td>
<td>0.7889</td>
<td>0.5053</td>
<td>0.006566</td>
</tr>
<tr>
<td>Precip of Wettest Quarter</td>
<td>-0.01466</td>
<td>0.8176</td>
<td>0.4867</td>
<td>-0.002061</td>
</tr>
</tbody>
</table>
Supplemental Figure 4.1 Schematic of PCR validation assay for large deletions and an example assay gel
Supplemental Figure 4.2 The first two principal components acquired using only deletion polymorphisms and colored by the population where they were originally collected.
Supplemental Figure 4.3 The first two principal components acquired using only SNPs and colored by the population where they were originally collected.
**Supplemental Figure 4.4** Kernel density estimates of the fixation index for SNPs, small deletions, and small insertions discovered using mpileup. Positive values indicate an excess of homozygosity relative to that expected under Hardy-Weinberg Equilibrium.
Supplemental Figure 4.5 Density of small deletions within and around recombination hotspots. The green dot gives the mean density of small deletions in deletions per base pair for 606 recombination hotspots. Blue dots give the mean density for 2500bp non-overlapping bins extending away from the boundaries of the hotspot region. The highlighted blue range indicates a 95% CI for the expected mean indel density in recombination hotspots based on permutations throughout the genome.
Supplemental Figure 4.6 Gene expression for genes with CDS impacted by deletions. (A) Mean expression values for genes impacted by deletions (black) vs. the expected range of mean values (grey). (B) Mean Pearson coexpression between Salicoid duplicates impacted by deletions (black) vs. expected. Error bars indicate 95% confidence intervals. (*** p < 0.001, ** p < 0.01, * p < 0.05; Two-tailed)
Supplemental Figure 4.7 Observed (black) and expected values for Salicoid frameshift deletions
Supplemental Figure 4.8 Observed (black) and expected values for Salicoid non-frameshift deletions
Supplemental Figure 4.9 Observed (black) and expected values for Salicoid large deletions
Supplemental Figure 4.10 Observed (black) and expected values for non-Salicoid frameshift deletions
Supplemental Figure 4.11 Observed (black) and expected values for non-Salicoid non-frameshift deletions
Supplemental Figure 4.12 Observed (black) and expected values for non-Salicoid large deletions
Supplemental Figure 4.13 Positions of small non-frameshift deletions within the largest group of bHLH transcription factors. Horizontal lines indicate sequences within a multiple sequence alignment, while red triangles indicate the positions of the deletions. Sequence conservation at each position is given by the vertical lines at the top, and the DNA binding domain is highlighted in blue.
Supplemental Figure 4.14 Positions of small frameshift deletions in the largest group of WRKY transcription factors. WRKY DNA binding domains are highlighted in blue.
Supplemental Figure 4.15 Scree plot of the principal components of environmental variation
**Supplemental Figure 4.16** Maps of the top 4 PCs of environmental variation. Red hues indicate higher values. Black dots signify sampling locations for the trees used in the association analysis.
Supplemental Figure 4.17 The number of unique genes implicated in associations with the first 4 PCs of environmental variation when performed on SNPs only and deletions only. Genes were deemed implicated if significantly associated markers resided within the genic regions or within the integenic region flanking them.
Supplemental Figure 4.18 Deletion association results for the environmental PC1. The horizontal lines indicate the 95% significance level based on permutation.
Supplemental Figure 4.19 Deletion association results for the environmental PC2.
Supplemental Figure 4.20 Deletion association results for the environmental PC3.
CHAPTER V

GENERAL CONCLUSIONS
On July 22, 1879, 20 years after the publication of On the Origin of Species, Charles Darwin described the rapid appearance of extant angiosperm lineages in the fossil record as an “abominable mystery” in a letter to J.D. Hooker (Darwin et al., 1903). Although the diversification of angiosperms is now thought to have occurred over a period of more than 45 million years (Coiffard et al., 2012), the origins of the angiosperms’ terrestrial dominance continue to provide an awe-inspiring puzzle. The recent advent of affordable whole genome sequencing has finally begun to shed light on the genetic basis of such a successful radiation, casting whole genome duplications in a central role (Jiao et al., 2011). In particular, the repeated retention of floral developmental genes following two WGDs at the base of seed plants and angiosperms may have equipped primordial flowering plants with the toolkit to for subsequent reproductive diversification (Jiao et al., 2011). The principal of uniformitarianism, which originally informed the development of Darwin’s theory, implies that WGDs continue to operate as a diversifying force today. Indeed, studies of several angiosperm lineages with more recent WGDs emphasize the role of divergently functioning WGD duplicates in diverse adaptations (Duarte et al., 2006; Rensing et al., 2008; Hsu et al., 2011). Therefore, the study of the evolution of paleopolyploid genomes is vital for understanding both historical and ongoing processes of genomic diversification.

Over the last 10 years, several studies of paleopolyploid genomes have invoked the role of dosage balance between interacting proteins to explain the biased retention of putatively dosage-sensitive genes such as transcription factors and members of conserved signal transduction pathways (Davis and Petrov, 2005; Thomas et al., 2006; Hughes and Liberles, 2008; Buggs et al., 2012). In Populus trichocarpa, I found a similarly biased pattern of duplicate retention (Chapter 2), with those same classes of genes highly enriched following the Salicoid
WGD but depleted among tandem duplicates. Furthermore, genes retained from the WGD have expression across greater numbers of tissues than tandem duplicates but tend to have slower rates of evolution. This pattern is consistent with the implications of the gene balance hypothesis, which predict that the dosage constraints enforcing purifying selection for retention of a given dosage sensitive gene after a balanced WGD would also promote the loss of the gene’s nascent small-scale duplications within a population (Veitia et al., 2008).

The results of chapter 2 also suggest that not all retained duplicates from the Salicoid WGD have the same constraints, as expressional divergence varies substantially between duplicate pairs. Specifically, while Salicoid duplicates overall tend to have more conserved expression than expected under a random pattern of divergence, the extent of conservation varies by breadth of expression and gene function. Duplicate pairs that, together, are expressed in a wider array of tissues also tend to have more conserved expression patterns. Additionally, transcription factors are enriched among genes with conserved expression, while transmembrane proteins and receptors are enriched among duplicates exhibiting expressional dissimilarity concordant with a model of random divergence, such as would occur under subfunctionalization (Force et al., 1999). This suggests that purifying selection for the maintenance of ancestral expression patterns varies considerably among retained duplicates, with transcription factors and paralogs stemming from putatively ubiquitously expressed genes under particularly high constraint but other genes, such as receptors, under lower constraint. This lack of constraint for maintenance of ancestral expression patterns in some genes also implies that they may be more prone to nonfunctionalization in an ongoing fractionation process. Indeed, I provide evidence for this in chapter 4, where I show that Salicoid duplicates pairs affected by segregating frameshift polymorphisms also have significantly reduced coexpression.
Although the pattern of WGD duplicate retention among gene categories provides support for the role of dosage constraint, it does not directly address the relationship between connectivity in defined biological networks with the probability of duplicate retention. This prompted my development of a statistical learning method for the prediction of physical interactions (chapter 3). The resulting predicted interactions are not only biologically plausible – as assessed by their functional similarity, higher coexpression, and involvement in related metabolic pathways – but also provide further insight into duplicate retention. They illustrate the importance of relative dosage during the fractionation process, as the probability of retention related to the interaction of the absolute connectivity and the frequency of interactors with retained duplicates in several independent WGDs. Moreover, the main effect of connectivity has a negative relationship with retention probability once this interaction is taken into account, as would be expected if selection for maintenance of stoichiometric balance were acting on a subnetwork with large numbers of singletons. Altogether, the results of this analysis further implicate gene balance as a prominent determinant of fitness in the paleopolyploid genome and suggest that its power is arbitrated by connectivity.

While the results of chapters 2 and 3 heavily suggest the involvement of dosage balance in past fractionation events, those of chapter 4 reveal its ongoing involvement in the selective regime mediating continued duplicate retention. I find that the patterns of functional enrichment for segregating, putatively nonfunctionalizing, indel polymorphisms parallel the patterns of long-term duplicate retention observed in chapter 2, with the abrogation of transcription factor function heavily disfavored and the loss of duplicate catalytic function comparatively unencumbered by purifying selection. Interestingly, the enrichment of small non-frameshifting indels among transcription factors suggests that these may provide a basis for diversification of
transcriptional networks following WGD, highlighting the potential for generating adaptive novelty in addition to the more general contribution of indels to genetic load. Again, I find that the network context of genes is associated with the probability of nonfunctionalization, both with respect to predicted PPI networks and metabolic networks. In the case of the predicted PPI networks, the effect of connectivity is negatively associated with the probability of a segregating nonfunctionalizing polymorphism for Salicoid duplicates and Salicoid singletons alike, suggesting overall increased purifying selection for more connected genes. However, the results also suggest an ongoing role for relative dosage, as Salicoid catalytic genes with segregating nonfunctionalizing polymorphisms also have lower relative frequencies of duplicated neighbors.

Beyond revealing the stability of biases underlying duplicate gene retention, the large and small indels analyzed in chapter 4 add to our knowledge of how genetic diversity is generated and provide insight into the genetic basis of environmental adaptation. In contrast to a previous report in humans (Montgomery et al., 2013), I find that small indel polymorphisms are enriched in and around recombination hotspots. I also provide support for a previously proposed positive relationship between indel and SNP polymorphisms (Tian et al., 2008; Hollister et al., 2010). These results raise the intriguing possibility that recombination adds substantial novel genetic variation in addition to breaking down associations between pre-existing variants. The relative magnitudes of these two consequences of crossing over and their theoretical implications within the population are worthy targets of future study. Several large and small deletions are also significantly associated with the primary components of environmental variation, implicating loci that are not significantly associated with any SNPs but have similar biological functions to loci that are solely associated with SNPs. Therefore, the addition of indel variation to the
reertoire of polymorphisms assayed during association studies may significantly improve the sensitivity of GWAS, especially when effect sizes are small and linkage disequilibrium is low.

The prominence of WGDs throughout the history of flowering plant evolution suggests that the processes underlying retention and loss of duplicated genes have shaped the phenotypic diversity that originally captivated Darwin 150 years prior. Throughout this work, I have shown that the functional biases among WGD duplicates in other sequenced plant genomes are present in *P. trichocarpa* and, to a large extent, manifest in a manner consistent with past and ongoing selection for the maintenance of dosage balance. Future studies of evolution following genome duplication and adaptive variation within the paleopolyploid genome would therefore likely benefit from increased attention to biological networks and an examination of how current evolutionary forces relate to the signatures of past duplicate retention and loss.
References


Thomas, B. C., Pedersen, B., and Freeling, M., 2006. Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, **16**:934–946.
Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R.,