



# **The impact of recurrent polyploidy on photosynthesis in Glycine**

by Jeremy Eugene Coate

---

This thesis/dissertation document has been electronically approved by the following individuals:

Doyle, Jeffrey J (Chairperson)

Owens, Thomas G (Minor Member)

Van Wijk, Klaas (Minor Member)

Kresovich, Stephen (Field Appointed Member Exam)

THE IMPACT OF RECURRENT POLYPLOIDY ON PHOTOSYNTHESIS IN  
*GLYCINE*

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jeremy Eugene Coate

August 2010

© 2010 Jeremy Eugene Coate

THE IMPACT OF RECURRENT POLYPLOIDY ON PHOTOSYNTHESIS  
IN *GLYCINE*

Jeremy Eugene Coate, Ph. D.

Cornell University 2010

Polyploidy (whole genome duplication) is a ubiquitous feature of flowering plants that produces pronounced effects on phenotype, yet the underlying genetic mechanisms for these effects are mostly unknown. Photosynthesis is one of the most striking examples of how polyploids can differ dramatically from their diploid progenitors. We utilized the recurrent history of genome duplication within the genus *Glycine* to characterize the genetic and genomic mechanisms through which polyploidy impacts photosynthesis. Using genomic resources for the cultivated soybean (*G. max*), we examined how photosynthetic gene families have been shaped by ancient polyploidy events, as well as by non-polyploid (e.g., tandem) duplications, and found fundamental differences in the patterns of retention and loss for gene families encoding subunits of photosystems I and II compared to the Calvin cycle. We further showed that equivalent patterns emerge in two other paleopolyploids, *Medicago truncatula* and *Arabidopsis thaliana*. These differences suggest that photosystem gene families are dosage sensitive, whereas Calvin cycle gene families are not. We then devised an assay to quantify whole transcriptome size, as well as the expression of individual genes on a per-cell basis, enabling quantification of gene dosage responses and facilitating comparisons of gene expression across ploidy levels. Using this assay we demonstrated that the transcriptome of a recently formed *Glycine* allotetraploid (*G. dolichocarpa*) is ca. 1.4-fold larger than those of its diploid

progenitors (*G. tomentella* and *G. syndetika*), and that most genes exhibit partial dosage compensation in the tetraploid. Because polyploidy is associated with enhanced stress tolerances, we also examined the effects of polyploidy on nonphotochemical quenching (NPQ), a set of mechanisms to protect the photosynthetic machinery under excess light stress. We showed that the *G. dolichocarpa* tetraploid has enhanced NPQ capacity under excess light compared to its diploid progenitors. Transcript profiling revealed that the tetraploid over-expresses two classes of galactolipid synthase genes, which results in altered leaf lipid profiles that may explain the differences in NPQ.

## BIOGRAPHICAL SKETCH

Jeremy Coate received a bachelor's degree in biology from Reed College in Portland, Oregon in 1992. He then spent two years working as a Peace Corps Volunteer in The Gambia, West Africa, developing agroforestry systems with local farmers. Jeremy subsequently earned a master's degree in forestry from Oregon State University in 1999. At Oregon State, Jeremy worked for the Tree Genetic Engineering Research Cooperative where he developed an interest in using molecular techniques to study plant biology. He then spent four years working as a molecular biologist at Exelixis Plant Sciences in Portland, Oregon, identifying genes controlling various traits of agronomic interest. Jeremy started his graduate work in the Department of Plant Biology at Cornell in August of 2004.

## ACKNOWLEDGMENTS

First and foremost, I thank my advisor, Jeff Doyle, for his guidance over the past six years. It has been an honor and a pleasure to work with someone who is not only an excellent scientist, but also a kind and dedicated mentor.

I thank past and present members of the Doyle lab, all of whom have provided useful advice and valuable insights at various times in my graduate career. Jane Doyle and Sue Sherman-Broyles (a.k.a., “the lab moms”) deserve special mention for making the lab run so smoothly in addition to carrying out their own research projects. I also thank Sue for the many helpful, thought-provoking, reassuring, stress-relieving, and/or plain silly conversations we shared while carpooling to and from Cortland (and apologize for those times I lost my keys or ran out of gas).

I thank my other committee members, Klaas van Wijk and Tom Owens, for many thoughtful suggestions as well as for their critical reading of this dissertation. Tom also devoted countless hours of his time teaching me how to collect and interpret photosynthetic data, as well as helping to write the grants that have funded my research, and it was an honor working with such an excellent physiologist and teacher.

I have had the privilege of collaborating with several excellent scientists at other institutions who deserve mention here. Greg May at the National Center for Genome Resources in New Mexico provided the opportunity to apply next generation sequencing technology to my research, fundamentally changing my project and opening a whole new world of possibilities that I am still just beginning to explore. Andrew Farmer, also at NCGR, has been incredibly helpful in making sense of our RNA-Seq data. Jessica Schlueter at UNC Charlotte was instrumental in the bioinformatic analyses described in chapter one. Thanks to and Steven Cannon at Iowa State and Gary Stacey at the University of Missouri for letting me use their soybean

RNA-Seq data. Eric Marechal at CEA Grenoble, France was kind enough to perform the lipid profiling analyses described in chapter four, and was very helpful in interpreting the data.

I gratefully acknowledge funding from three National Science Foundation grants (DEB-0709965, IOS-0744306, and IOS-0939423), as well as from the Department of Plant Biology at Cornell.

Finally, I thank my family. My parents, Chuck and Joyce, are a constant source of inspiration and support that I have relied upon heavily. My daughter, Juniper, has made the past year better than all those that preceded it. My wife, Breanne, is also my best friend, and is more important to me than she could ever know.

## TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS	vi
LIST OF FIGURES	vii
LIST OF TABLES	ix
CHAPTER 1. Introduction	1
CHAPTER 2. Comparative evolution of photosynthetic genes in response to polyploid and non-polyploid duplication	17
CHAPTER 3. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid	66
CHAPTER 4. Enhanced photoprotection in a recent allotetraploid correlates with high levels of galactolipid synthase gene expression and modified galactolipid profiles	102

## LIST OF FIGURES

FIGURE 1.1. The <i>Glycine</i> subgenus <i>Glycine</i> polyploid complex	3
FIGURE 1.2. Anatomical and morphological differences between a T2 allotetraploid and representatives of its diploid progenitor species (D3 and D4)	4
FIGURE 2.1. Estimated timing of genome duplication events in <i>Arabidopsis</i> , <i>Glycine</i> and <i>Medicago</i>	20
FIGURE 2.2. An example of percent retention and percent expansion calculations	23
FIGURE 2.3. Percent retention of homoeologues, given by gene family and species	24
FIGURE 2.4. Observed retention rates for photosynthetic genes following polyploidy in <i>Glycine</i>	25
FIGURE 2.5. Percent expansion of gene families by species	27
FIGURE 2.6. Duplicate retention by duplication category and functional group in <i>Glycine</i> , <i>Medicago</i> and <i>Arabidopsis</i>	29
FIGURE 2.7. Fractions of homoeologue pairs from the most recent polyploidy event ( $\alpha$ or A) exhibiting evidence for functional divergence	37
FIGURE 2.8. Heat maps of expression correlation coefficients ( $r$ ) within photosynthetic functional groups in <i>Arabidopsis</i>	40
FIGURE 3.1. A comparison of transcriptome-normalized expression data vs. genome-normalized expression data	69
FIGURE 3.2. qRT-PCR based estimates of transcripts per genome and RNA-Seq based estimates of transcripts per transcriptome	80
FIGURE 3.3. T2 transcriptome size relative to the transcriptomes of its diploid progenitors	83
FIGURE 3.4. Genome-wide distribution of gene dosage responses and homoeologue silencing in the T2 allotetraploid	85

FIGURE 4.1. NPQ response curves	110
FIGURE 4.2. Contributions of energy-dependent quenching (qE), protective photoinhibition (qI <sub>P</sub> ) and damage-induced photoinhibition (qI <sub>D</sub> ) to total NPQ	111
FIGURE 4.3. qRT-PCR validation of RNA-Seq expression estimates for five genes	113
FIGURE 4.4. RNA-Seq based estimates of combined expression for gene families encoding PsbS and xanthophyll cycle enzymes	114
FIGURE 4.5. RNA-Seq based estimates of combined expression for gene families encoding photosystem II-associated light harvesting proteins	115
FIGURE 4.6. RNA-Seq based estimates of combined expression for gene families encoding photosystem II (PSII) subunits	116
FIGURE 4.7. RNA-Seq and qRT-PCR estimates of combined MGD gene expression	117
FIGURE 4.8. RNA-Seq and semi-quantitative RT-PCR estimates of combined DGD gene expression	119
FIGURE 4.9. Leaf galactolipid profiles under LL and EL	120

## LIST OF TABLES

TABLE 2.1. Photosynthetic gene families, by functional group, and their sizes in <i>Arabidopsis</i> , <i>Glycine</i> and <i>Medicago</i>	22
TABLE 2.2. Average percent retention and percent expansion following polyploidy events	28
TABLE 2.3. Sliding window estimates of selection ( $\omega$ ) by functional group for the most recent polyploidy events ( $\alpha$ and $A$ ) in <i>Arabidopsis</i> and <i>Glycine</i>	36
TABLE 2.4. Degree of expression correlation between duplicate gene pairs by duplication type and functional group in <i>Arabidopsis</i> and <i>Glycine</i>	39
TABLE 3.1. Genes and gene families for which expression was analyzed by genome-normalized qRT-PCR	74

# CHAPTER 1

## INTRODUCTION

### *Polyploidy and photosynthesis*

Polyploidy (whole genome duplication) is ubiquitous in flowering plants, and likely played a central role in their origin and radiation (De Bodt et al. 2005).

Extensive synteny within sequenced genomes provides evidence for a hexaploidy event in the common ancestor of the two largest clades of eudicots (Tang et al. 2008), and chromosomal diploids such as *Arabidopsis*, rice, and poplar show evidence of additional polyploid duplications (Bowers et al. 2003; Sterck et al. 2005; Zhang et al. 2005; Tuskan et al. 2006). Thus most, if not all, flowering plants are paleopolyploids, and it is clear from these species and from other, less fully-characterized taxa (Blanc and Wolfe 2004a; Cui et al. 2006; Pfeil et al. 2005; Schlueter et al. 2004; Schranz and Mitchell-Olds 2006; Town et al. 2006; Barker et al. 2008) that flowering plant genomes comprise nested sets of duplications.

In addition to the prevalence of polyploid lineages, considerable evidence suggests that genome doubling is adaptive. Polyploids frequently have broader geographic ranges than their diploid progenitors (Doyle et al. 2004, Otto and Whitton 2000), and appear to be more successful in extreme or stressful habitats (Ehrendorfer 1980, Lewis 1980, Otto and Whitton 2000). Recent work suggests that polyploid plant lineages were more likely to survive the Cretaceous–Tertiary mass extinction than were diploid lineages (Fawcett et al. 2009).

The success of polyploids is perhaps not surprising when considering that polyploids can differ dramatically and in evolutionarily significant ways from their diploid progenitors. Among numerous other examples, polyploid wheat exhibits

greater water and nitrogen use efficiencies, as well as grain yields, than diploid wheat (Huang et al. 2007), and synthetic *Arabidopsis* allotetraploids show increased growth vigor (Ni et al. 2009).

Photosynthesis plays a fundamental role in plant fitness, and is perhaps one of the most striking and well documented examples of how polyploids can differ dramatically from their diploid progenitors across a broad range of taxa (Warner and Edwards 1993). Though most photosynthetic parameters show variable responses, polyploids typically exhibit more chloroplasts per mesophyll cell, and higher photosynthetic rates per cell (Warner and Edwards 1993). The causes of these biochemical changes at the level of underlying genes are largely unknown.

The following chapters have in common a focus on the evolutionary consequences of polyploidy, with particular emphasis on how genome duplication affects photosynthesis. Because polyploidy is associated with greater stress tolerances, we also examined the effects of polyploidy on photoprotection (Chapter Four), a set of mechanisms to protect the photosynthetic machinery under excess light stress. Most of the existing studies of photosynthesis in polyploids are physiological in focus, and predate the genomics era (e.g., Warner and Edwards 1993). We have utilized existing genomics data, as well as newer genomics technology (Next Generation Sequencing), in order to explore the relationship between physiological and genetic/genomic responses to genome duplication.

### ***The genus *Glycine* as a model system for studying polyploidy***

The legume genus *Glycine* is divided into two subgenera, *Soja* (which includes the cultivated soybean, *G. max*) and *Glycine*, a group of around 25 perennial species centered in Australia. The genus experienced two rounds of polyploidy (ca. 54 MYA and 13 MYA) prior to divergence of the two subgenera (Schlueter et al. 2004,

Schmutz et al. 2010). A third burst of genome duplication occurred in subgenus *Glycine* within the last 100,000 years, producing an extensive and well-studied allopolyploid complex (reviewed in Doyle et al., 2004) (Figure 1.1). This complex includes eight allopolyploid ( $2n = 78, 80$ ) species derived from various combinations of diploid ( $2n = 38, 40$ ) genomes (Figure 1.1). Thus, *Glycine* is an attractive system for studying patterns of evolution following polyploidy, particularly in light of the fact that the soybean genome sequence was recently completed (Schmutz et al. 2010).

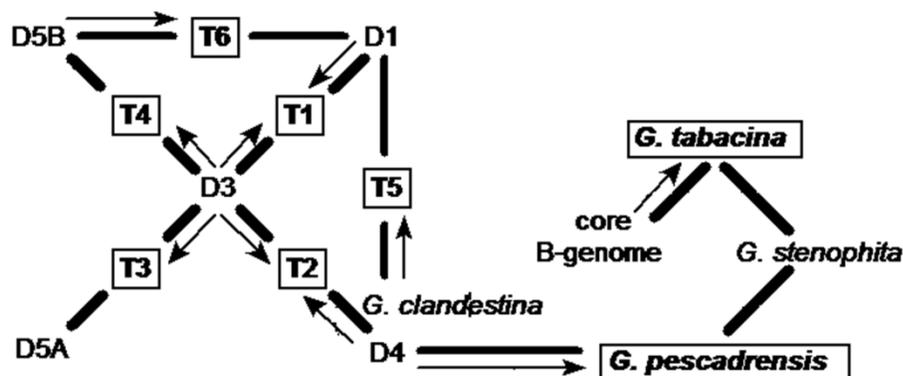


Figure 1.1. The *Glycine* subgenus *Glycine* polyloid complex. Taxa indicated by letters (e.g., D1, T3) are all classified as *G. tomentella* "races" (= reproductively isolated species not yet recognized taxonomically) at either the diploid ("D") or tetraploid ("T") level. Polyploid taxa are boxed (e.g., T1-T6, *G. pescadrensis*), and are connected to their diploid progenitors by heavy lines; arrows indicate chloroplast donors (e.g., *G. clandestina* was the sole chloroplast donor to the T5 polyploid, whereas both D1 and D3 diploids contributed chloroplast genomes to the T1 polyploid). (Modified from Doyle et al. 2004.)

Members of the subgenus *Glycine* polyloid complex individually illustrate many evolutionary features of allopolyploidy and provide strong genetic, evolutionary and ecological contrasts with one another. All diploids are confined to Australia, but some polyploids are extensive colonizers, ranging far beyond their diploid progenitors

to islands of the Pacific Ocean. Polyploids also exhibit a range of anatomical and morphological differences when compared to their diploid progenitors (Figure 1.2).

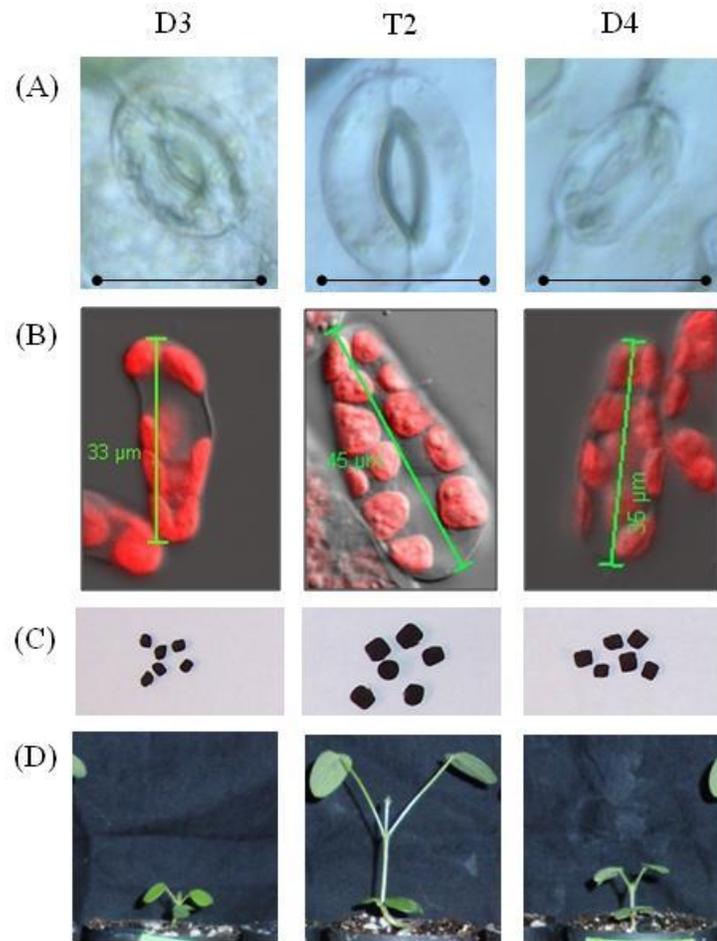


Figure 1.2. Anatomical and morphological differences between a T2 allotetraploid and representatives of its diploid progenitor species (D3 and D4). (A) Guard cells (bars = 20µm); (B) Palisade cells; (C) seeds; (D) seedlings approximately one week post-germination.

In addition, all of these allopolyploid species are estimated to have originated within the last 50,000 to 100,000 years, which corresponds to major environmental changes in Australia (Hope, 1994). The last ice age began approximately 60,000 years ago, resulting in significant cooling and drying. At approximately the same time,

humans first colonized the continent, and are thought to have rapidly altered the vegetation by burning and over-hunting (Hope 1994, Brooks and Bowman 2002, Hudjashov et al. 2007). Combined, these events presumably opened new niches characterized by increased exposure to direct sunlight, and decreased moisture and temperature – conditions likely to induce photoinhibition. We speculated that allopolyploids could have succeeded during this time – as at no other time in the history of the subgenus – in part because allopolyploidization generated enhanced photoprotective capabilities.

### ***Polyploidy and the evolution of photosynthesis in Glycine***

Each of the following three chapters has utilized the history of polyploidy in *Glycine* to explore, at various levels, the effects of genome duplication on photosynthesis. Briefly, utilizing the genomic resources of soybean (*G. max*), Chapter Two examines how photosynthetic gene families have evolved following the two *Glycine* paleopolyploidy events (ca. 54 MYA and 5-13 MYA), as well as in response to non-polyploid (e.g., tandem) duplications, and provides a baseline for understanding photosynthesis in recently formed polyploids. Chapters Three and Four focus on physiological and genomic responses following recent genome duplication by comparing a subgenus *Glycine* neopolyploid (T2; Figure 1.1) with its diploid progenitors (D3 and D4).

*Chapter 2. Comparative evolution of photosynthetic genes in response to polyploid and non-polyploid duplication.* The objective of this study was to characterize the effects of polyploidy, as well as non-polyploid duplications, on the network of functionally interrelated genes underlying photosynthesis. Genomic studies have begun to elucidate distinct patterns of gene retention and loss, correlating with functional classification, for both polyploid and non-polyploid duplications (Blanc and

Wolfe 2004b; Freeling and Thomas 2006; Maere et al. 2005; Seoighe and Gehring 2004). Transcription factors and kinases, for example, exhibit high retention rates after polyploidy, and low retention rates following non-polyploid duplication (Blanc and Wolfe 2004b; Maere et al. 2005). Conversely, several gene ontology (GO) categories, including DNA metabolism, show the opposite pattern. Though taxonomic sampling is limited to date, in many cases these patterns appear to hold across a range of species.

To date, such patterns have been identified in the context of protein domains (Paterson et al. 2006), GO categories (Blanc and Wolfe, 2004b; Freeling and Thomas 2006; Maere et al. 2005; Seoighe and Gehring 2004), and co-expressed networks (Blanc and Wolfe 2004b), but little is known of the effect of polyploidy on the gene networks that underlie key physiological or developmental processes. Because photosynthesis is known to be affected by polyploidy (Warner and Edwards 1993), we investigated how whole genome duplication, as well as single gene duplications, has shaped the structure of photosynthetic gene families.

This analysis revealed distinct patterns of duplicate retention for photosystem gene families compared to the Calvin cycle. Overall, these patterns suggest that the photosystems are dosage sensitive (changes in the amounts of some subunits but not others impairs complex assembly and/or function, and are deleterious), whereas the Calvin cycle is not. Duplicates of Calvin cycle genes, in contrast, exhibit more capacity for functional differentiation than do photosystem genes.

*Chapter 3. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid.* This chapter describes a novel method of normalizing qRT-PCR experiments that makes it possible to measure relative expression per cell across species or ploidy levels. Coupling this approach with transcript profiling data enables estimates of relative transcriptome size (total number of mRNA transcripts per cell).

All widely used methods of quantifying expression yield transcriptome-normalized expression values. Without knowing the sizes of the transcriptomes being compared, it is not possible to infer expression level per cell. This is a severe limitation in comparisons across ploidy levels, where genome-wide differences in gene dosage likely affect overall transcriptome size. Even for many comparisons at the diploid level (e.g., comparing leaf tissue and root tissue), transcriptomes may well differ in size. Our method to measure gene expression per cell, as well as overall differences in transcriptome size, will therefore added valuable information to traditional transcript profiling experiments, particularly those involving polyploidy.

Additionally, the patterns of duplicate retention observed in Chapter Two, as well as in other studies (Blanc and Wolfe 2004b; Freeling and Thomas 2006; Maere et al. 2005; Seoighe and Gehring 2004), support the hypothesis that gene family evolution is, in some cases, constrained by dosage sensitivity (i.e., there is selection to maintain balance in copy number among gene families encoding subunits of protein complexes). Dosage sensitivity correlates with the extent to which a gene's product forms protein-protein interactions; highly "connected" genes tend to retain polyploid duplicates and eliminate non-polyploid duplicates (Thomas et al. 2006). There are, however, numerous exceptions. Genes that appear to meet the criteria of being connected, but do not follow the predictions of the balance hypothesis, may represent genes for which transcript abundance is readily decoupled from gene dosage (Veitia et al. 2008; Edger and Pires 2009). Consequently, the method described in this chapter to estimate dosage responses across the genome, will help to test and refine the balance hypothesis.

In this chapter we coupled our novel genome-normalized qRT-PCR assay with transcript profiling data (RNA-Seq) in order to obtain the first estimate of the size of a tetraploid transcriptome relative to its diploid progenitor transcriptomes. We showed

that the T2 leaf transcriptome is 1.4-fold larger than the midparent diploid transcriptome (i.e., has 1.4-fold more mRNA molecules per cell), and that the majority of genes exhibit partial dosage compensation. We further showed that rates of homoeologue silencing differ significantly by dosage response.

*Chapter 4. Enhanced photoprotection in a recent allotetraploid is associated with modified galactolipid profiles.* Though numerous studies have shown pronounced effects of polyploidy on photosynthesis, photoprotection has never previously been examined in the context of polyploidy. Here we showed that the T2 allotetraploid has enhanced photoprotective capacity (NPQ) compared to its diploid progenitors (D3 and D4). We then provided evidence to suggest that this increase is the result of differences in leaf galactolipid composition. Based on transcript profiling data, we showed that two galactolipid synthase gene families are up-regulated in T2 compared to D3 or D4. We next showed that this resulted in altered galactolipid profiles, including higher levels of the galactolipid, digalactosyldiacylglycerol (DGDG). Galactolipids are the major lipid constituents of thylakoid membranes (Benning and Ohta 2005), and are intimately associated with photosystem II (Loll et al. 2005). Mutants deficient in MGDG or DGDG have been shown previously to be more susceptible to photoinhibition (Aronsson et al. 2008, Holzl et al. 2009).

Several strategies for NPQ are evolutionarily conserved throughout the plant kingdom (Avenso et al., 2004; Horton and Ruban, 2005), and are essential for plant fitness in the field (e.g., Kulheim et al., 2002). Our data suggest that adaptations that increased the capacity of the photoprotective apparatus or its flexibility to adjust to different light environments could potentially have contributed to the success and expanded geographical ranges of some polyploids, including the *Glycine* T2 allopolyploid studied here.

### *Current and Future directions*

In addition to the existing RNA-Seq data for D3, D4 and T2 used in Chapters Three and Four, we are currently generating transcript profiling data for two more *Glycine* neopolyploid species (T1 and T5; Figure 1.1) and two more diploid species (A and D1). Once complete, we will have biological replicates for four accessions from each of seven species (three allopolyploids and their four diploid progenitors; Figure 1.1) under both limiting and excess light. Two synthetic allopolyploids involving the same diploid genome combinations as T5 will be included. This dataset will enable us to look for patterns in transcriptomic responses to allopolyploidy, and to further explore several aspects of the work described in this dissertation:

1. *Gene dosage responses*: As described in Chapter Three, RNA-Seq expression data will be coupled with genome-normalized qRT-PCR to estimate transcriptome sizes, and dosage responses of individual genes. By mapping dosage responses to gene ontology (GO) classifications, we will determine if GO categories that are preferentially retained following whole genome duplication exhibit specific dosage responses in the period of time immediately following polyploidy. Preliminary results indicate that GO categories showing preferential retention following genome duplication (e.g., kinases and transcription factors) are enriched for genes exhibiting a 1:1 dosage effect (doubling of transcript abundance with a doubling of gene copy number). We will perform similar analyses on photosynthetic gene families to see if the retention patterns described in Chapter Two correlate with particular dosage responses in the neopolyploids.
2. *Expression responses to excess light*: We have also quantified NPQ capacity in each of the accessions for which RNA-Seq data are being generated.

Preliminary data indicate that individual accessions of all three polyploid species achieve greater NPQ capacities than their diploid progenitors. By coupling these phenotypic data with transcript profiling data we can search for patterns in expression responses that may underlie the enhanced NPQ phenotype. For example, do all polyploids with high NPQ capacities exhibit up-regulation of MGD and DGD genes?

We are also quantifying photosynthetic capacity (linear electron transport rate) in the same species and accessions. Photosynthesis and NPQ represent alternative pathways for the dissipation of absorbed light energy. NPQ mechanisms appear to be activated only to the extent that light absorption exceeds the energy-utilizing capacity of photochemical quenching pathways (Melkonian et al., 2004). Thus, the selective pressures on NPQ capacity are presumably modulated by the capacity for photochemical quenching. By quantifying photosynthetic capacity, we will gain a more complete understanding of the evolution of NPQ in *Glycine* allopolyploids. These studies will allow us to assess whether there are patterns in the responses of photosynthesis to allopolyploidy across independent allopolyploid species. They will also give distinct and complementary perspectives on photosynthesis. Linear electron transport capacity reflects the functioning of the photosynthetic apparatus as a whole, integrating the performance of the full suite of photosynthetic proteins, as well as downstream components (e.g., sink strength) and alternative electron sinks (e.g., photorespiration). In contrast, responses to excess light occur primarily in PSII (Owens, 1996). Preliminary data indicate that some T2 polyploid accessions have elevated photosynthetic capacities relative to their diploid progenitors. Intriguingly, T2 accessions with D3 plastids (D3 maternal parent) have higher photosynthetic capacities than T2 accessions with D4 plastids (D4 maternal parent).

In order to better understand these photosynthetic phenotypes, we are also quantifying a number of anatomical properties of leaf palisade cells, including cell volume, number of chloroplasts per cell, and chloroplast volume per cell, as well as number of palisade cells per unit leaf area and chlorophyll content per unit leaf area. By coupling these data with per cell gene expression estimates (as described in Chapter Three) and NPQ capacity (Chapter Four), we hope to develop a comprehensive picture of the mechanisms by which allopolyploidy modifies photosynthesis and NPQ in the *Glycine* allopolyploid complex.

## REFERENCES

- Aronsson H, Schottler MA, Kelly AA, Sundqvist C, Dormann P, Karim S, Jarvis P. 2008. Monogalactosyldiacylglycerol deficiency in arabidopsis affects pigment composition in the prolamellar body and impairs thylakoid membrane energization and photoprotection in leaves. *Plant Physiol* 148:580-592.
- Avenson TJ, Cruz JA, Kramer DM. 2004. Modulation of energy-dependent quenching of excitons in antennae of higher plants. *Proc Natl Acad Sci* 101:5530-5535.
- Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore W, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* 25:2445-2455.
- Benning C, Ohta H. 2005. Three enzyme systems for galactoglycerolipid biosynthesis are coordinately regulated in plants. *J Biol Chem* 280:2397-2400.
- Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667-1678.
- Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. *Plant Cell* 16:1679-1691.
- Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433-438.
- Brook BW, Bowman DMJS. 2002. Explaining the Pleistocene megafaunal extinctions: models, chronologies, and assumptions. *Proc Natl Acad Sci* 99: 14624-14627.
- Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16:738-749.

- De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20:591-597.
- Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. 2004. Evolution of the perennial soybean polyploid complex (glycine subgenus glycine): A study of contrasts. *Biol J Linn Soc* 82:583-597.
- Edger P, Pires J. 2009. Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res* 17:699-717.
- Ehrendorfer F. 1980. Polyploidy and distribution. In: Lewis WH, editor. *Polyploidy: Biological Relevance*. NY: Plenum. p. 45-60.
- Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc Natl Acad Sci* 106:5737-5742.
- Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* 16:805-814.
- Holz G, Witt S, Gaude N, Melzer M, Schottler MA, Dormann P. 2009. The role of diglycosyl lipids in photosynthesis and membrane lipid homeostasis in arabidopsis. *Plant Physiol* 150:1147-1159.
- Hope GS. 1994. Quaternary vegetation. In: Hill RS, editor. *History of the Australian vegetation : Cretaceous to recent*. New York: Cambridge University Press, 1994. p. 368-389.
- Horton P, Ruban A. 2005. Molecular design of the photosystem II light-harvesting antenna: Photosynthesis and photoprotection. *J Exp Bot* 56:365-373.
- Huang M, Deng X, Zhao Y, Zhou S, Inanaga S, Yamada S, Tanaka K. 2007. Water and nutrient use efficiency in diploid, tetraploid and hexaploid wheats. *J Integr Plant Biol* 49:706-715.
- Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R et al. 2007. Revealing the prehistoric

- settlement of australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci* 104:8726-8730.
- Kulheim C, Agren J, Jansson S. 2002. Rapid regulation of light harvesting and plant fitness in the field. *Science* 297:91-93.
- Lewis WH. 1980. Polyploidy in species populations. In: Lewis WH, editor. *Polyploidy: Biological Relevance*. NY: Plenum. p. 103-144.
- Loll B, Kern J, Saenger W, Zouni A, Biesiadka J. 2005. Towards complete cofactor arrangement in the 3.0[thinsp]Å resolution structure of photosystem II. *Nature* 438:1040-1044.
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci* 102:5454-5459.
- Melkonian J, Owens TG, Wolfe DW. 2004. Gas exchange and co-regulation of photochemical and nonphotochemical quenching in bean during chilling at ambient and elevated carbon dioxide. *Photosynthesis Res* 79:71-82.
- Ni Z, Kim E, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ. 2009. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* 457:327-331.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* 34:401-437.
- Owens TG. 1996. Processing of excitation energy by antenna pigments. In: Baker NR, editor. *Photosynthesis and the environment*. London, U.K.: Kluwer Academic Publishers. p. 1-24.
- Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ. 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. *Syst Biol* 54:441-454.

- Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47:868-876.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178-183.
- Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 18:1152-1165.
- Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet* 20:461-464.
- Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van De Peer Y. 2005. EST data suggest that poplar is an ancient polyploid. *New Phytol* 167:165-170.
- Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18:1944-1954.
- Thomas BC, Pedersen B, 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.
- Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ et al. 2006. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* 18:1348-1359.
- Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (torr. & gray). *Science* 313:1596-1604.
- Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: Genomic, transcriptomic and proteomic effects. *Trends Genet* 24:390-397.

Warner DA, Edwards GE. 1993. Effects of polyploidy on photosynthesis.  
Photosynthesis Res 35:135-147.

## CHAPTER 2

# COMPARATIVE EVOLUTION OF PHOTOSYNTHETIC GENES IN RESPONSE TO POLYPLOID AND NON-POLYPLOID DUPLICATION

### ABSTRACT

The likelihood of duplicate gene retention following polyploidy varies by broadly defined functional properties (e.g., gene ontologies or PFAM domains), but little is known about the effects of whole genome duplication on gene networks related functionally by a common physiological process. Here, we examined the effects of both polyploid and non-polyploid duplications on genes encoding the major functional groups of photosynthesis (photosystem I, photosystem II, the light harvesting complex, and the Calvin cycle) in the cultivated soybean (*Glycine max*), which has experienced two rounds of whole genome duplication. Photosystem gene families exhibit retention patterns consistent with dosage sensitivity (preferential retention of polyploid duplicates and elimination of non-polyploid duplicates), whereas Calvin cycle and light harvesting complex gene families do not. We observed similar patterns in *Medicago truncatula*, which shared the older genome duplication with *Glycine* but has evolved independently for ca. 50 million years, and in *Arabidopsis thaliana*, which experienced two nested polyploidy events independent from the legume duplications. In both *Glycine* and *Arabidopsis*, Calvin cycle gene duplicates exhibit a greater capacity for functional differentiation than do duplicates within the photosystems, which likely explains the greater retention of very old duplicates and larger average gene family size for the Calvin cycle relative to the photosystems.

## INTRODUCTION

Polyploidy (whole genome duplication) has played an important role in the evolutionary history of angiosperms and has even been suggested to underlie their origin and radiation (De Bodt et al. 2005), as well as increasing the likelihood of surviving the Cretaceous–Tertiary extinction (Fawcett et al. 2009). Based solely on chromosome numbers, 30% or more of flowering plants are polyploid (Soltis et al. 2009), and it has been estimated that 15% of angiosperm speciation events involve polyploidy (Wood et al. 2009). Synteny data from sequenced genomes provide evidence for a hexaploidy event in the common ancestor of the two largest clades of eudicots (Tang et al. 2008), and chromosomal diploids such as *Arabidopsis*, rice, and poplar show evidence of additional, subsequent polyploid duplications (Bowers et al. 2003; Sterck et al. 2005; Zhang et al. 2005; Tuskan et al. 2006). Thus, the true percentage of flowering plant taxa that are paleopolyploids is certainly higher, and it is clear from these species and from other, less fully-characterized taxa (Blanc and Wolfe 2004a; Cui et al. 2006; Pfeil et al. 2005; Schlueter et al. 2004; Schranz and Mitchell-Olds 2006; Town et al. 2006; Barker et al. 2008) that flowering plant genomes comprise nested sets of duplications.

Much effort has been made to identify emergent effects of polyploidy--the universal "rules" by which polyploidy functions (Doyle et al. 2008). In *Arabidopsis*, genomic studies have begun to elucidate distinct patterns of gene retention and loss, correlating with functional classification, for both polyploid and non-polyploid (NP) duplications (Blanc and Wolfe 2004b; Freeling and Thomas 2006; Maere et al. 2005; Seoighe and Gehring 2004). Transcription factors and kinases, for example, exhibit high retention rates after polyploidy, and low retention rates following NP duplication (Blanc and Wolfe 2004b; Maere et al. 2005). Conversely, several gene ontology (GO) categories, including DNA metabolism, show the opposite pattern. Though taxonomic

sampling is limited to date, in many cases these patterns appear to hold across a range of species. Grouping genes by Pfam domains, Paterson et al. (2006) observed similar patterns across *Arabidopsis*, rice, yeast, and pufferfish (*Tetraodon*). Barker et al. (2008) found consistent patterns of retention and loss by GO class across all tribes of Compositae, which have evolved separately for >30 million years following a shared paleopolyploid duplication. It is noteworthy, however, that these patterns differ substantially from those observed in *Arabidopsis*, suggesting that, at least in some cases, such patterns are lineage specific (Barker et al. 2008).

Despite the considerable progress that has been made in elucidating patterns of duplicate gene retention following polyploidy, as well as mechanisms driving these patterns, little is yet known of the effect of polyploidy on the gene networks that underlie key physiological or developmental processes. The behavior of genes has been studied in the context of individual gene families (Adams and Wendel 2005), protein domains (Paterson et al. 2006), GO categories (Blanc and Wolfe, 2004b; Freeling and Thomas 2006; Maere et al. 2005; Seighe and Gehring 2004), and co-expressed networks (Blanc and Wolfe 2004b), but not in the framework of a physiological process.

The objective of this study was to characterize the effects of polyploidy, as well as non-polyploid duplications, on the network of functionally interrelated genes underlying photosynthesis, a key determinant of the ecological success and economic utility of plants. Photosynthesis is a prime example of how polyploids can differ phenotypically from their diploid progenitors. Polyploids consistently exhibit larger mesophyll cells with more chloroplasts and greater photosynthetic capacities per cell than their diploid progenitors (reviewed in Warner and Edwards 1993). The causes of these differences at the level of underlying genes are unknown.

The legume genus, *Glycine*, which includes the cultivated soybean (*G. max*), has a history of recurring polyploidy. In addition to two paleopolyploidy events in the lineage leading to soybean (Figure 2.1), the wild, perennial relatives of soybean underwent a burst of genome duplications within the last 100,000 years involving various combinations of extant diploid genomes (Doyle et al. 2004).

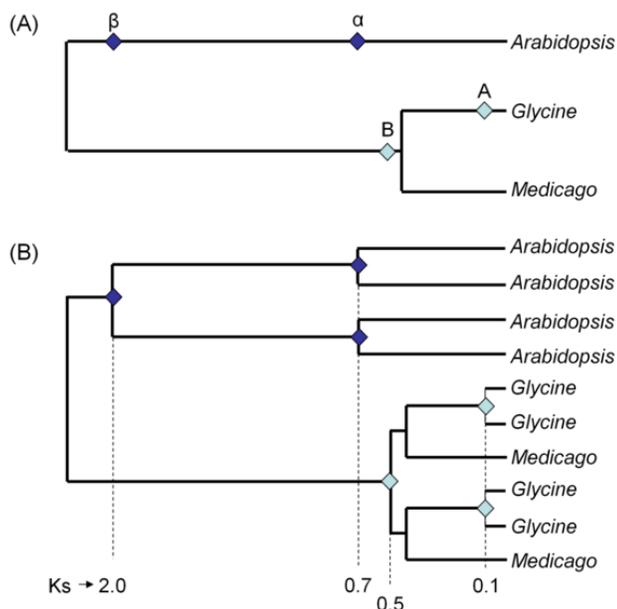


Figure 2.1. Estimated timing of genome duplication events in *Arabidopsis*, *Glycine* and *Medicago*. (A) A simplified species tree showing the relative maximum ages of genome duplication events (designated by diamonds), as estimated by homoeologue divergences. For duplication events in *Arabidopsis*, we follow the naming convention of Bowers et al. (2003), in which the most recent WGD is designated “α” and the older event “β.” The duplication events in *Glycine* are designated “A” and “B” to highlight the fact that they are distinct from (and more recent than) the *Arabidopsis* WGDs. Because *Medicago* shared the “B” duplication event with *Glycine*, we refer to this duplication as “B” in *Medicago* as well. The Phytozome gene clusters, “Rosid (pre-hexaploidy)” and “Rosid (post-hexaploidy)” predate the divergence of *Arabidopsis* and legumes; the “Eurosoid I” and “Legume” clusters fall between the common ancestor of *Arabidopsis* and legumes and the ancestor of *Medicago* and *Glycine*. (B) A gene tree showing the expected topology if all homoeologues have been retained in all three species. Diamonds represent the timing of homoeologue divergences, as indicated by synonymous substitutions per synonymous site ( $K_s$ ), which represent maximum ages for the polyploidy events (Doyle and Eagan 2010).

Thus, *Glycine* is an attractive system for studying patterns of genome evolution following polyploidy, particularly in light of the fact that the soybean genome sequence was recently completed (Schmutz et al. 2010). Here, we utilized the genomic resources of soybean to investigate how two nested rounds of whole genome duplication, as well as single gene duplications, have shaped the structure of photosynthetic gene families. Because the legume genus, *Medicago*, shared the oldest polyploidy event with *Glycine* (Figure 2.1; Pfeil et al. 2005), we performed similar analyses on the model species, *M. truncatula*, in order to determine the effects of a common polyploidy event in independently evolving lineages. Finally, we extended these analyses to the more distantly related eudicot model species, *Arabidopsis thaliana*, which has also experienced two well-characterized paleopolyploid events, in order to look for patterns emerging from independent sets of nested genome duplications. Thus, in total, we have analyzed photosynthetic gene family evolution across three plant species and four genome duplication events.

## RESULTS

Proteins of the Calvin cycle (CC), photosystem II (PSII), and photosystem I (PSI) are encoded by eleven, nine, and nine distinct nuclear gene families, respectively. Light harvesting complex (LHC) genes cluster into 12 distinct but distantly related nuclear gene families. Additional CC, PSII, and PSI proteins are encoded by the chloroplast, but because the focus of this study is on duplication events within the nuclear genome, plastid-encoded genes were not analyzed here. In all three species, CC gene families are the largest on average, and PSI gene families are the smallest (Table 2.1). *Glycine* has experienced the most recent polyploid duplication of the three species examined, and average photosynthetic gene family size is about twice as large in *Glycine* as in *Medicago* or *Arabidopsis* (Table 2.1).

Table 2.1. Photosynthetic gene families, by functional group, and their sizes in *Arabidopsis*, *Glycine* and *Medicago*.

Functional Group	Protein	Gene Family Size		
		<i>Glycine</i>	<i>Medicago</i>	<i>Arabidopsis</i>
CC	RbcS	10	6	4
	FBPase	10	3	3
	TPI	6	2	2
	PGK	4	2	3
	GAPDH	13	5	7
	FBA	14	4	8
	TKL	12	2	2
	RPE	4	3	3
	PRI	9	3	4
	PRK	2	1	1
	SBPase	3	1	1
	<b>Total/Avg</b>	<b>87/7.9</b>	<b>32/2.9</b>	<b>38/3.5</b>
PSII	PsbO	4	2	3
	PsbP	4	2	2
	PsbQ	3	1	2
	PsbR	2	1	1
	PsbS	3	1	1
	PsbTn	4	3	2
	PsbW	4	2	1
	PsbX	5	4	1
	PsbY	4	3	1
	<b>Total/Avg</b>	<b>33/3.7</b>	<b>19/2.1</b>	<b>14/1.6</b>
PSI	PsaD	2	1	2
	PsaE	4	2	2
	PsaF	1	1	1
	PsaG	2	1	1
	PsaH	4	1	2
	PsaK	2	1	1
	PsaL	2	2	1
	PsaN	2	1	1
	PsaO	2	1	1
	<b>Total/Avg</b>	<b>21/2.3</b>	<b>11/1.2</b>	<b>12/1.3</b>
LHC	LhcA1	2	1	1
	LhcA2	4	1	2
	LhcA3	2	1	1
	LhcA4	2	1	1
	LhcA5	2	1	1
	LhcA6	2	1	1
	LhcB1	8	6	5
	LhcB2	2	1	3
	LhcB3	2	1	1
	LhcB4	4	2	3
	LhcB5	4	1	2
	LhcB6	4	1	1
	<b>Total/Avg</b>	<b>38/3.2</b>	<b>18/1.5</b>	<b>22/1.8</b>
<b>Combined</b>	<b>Total/Avg</b>	<b>181/4.4</b>	<b>81/2.0</b>	<b>86/2.1</b>

We quantified the contributions of the various duplication events to each gene family using two parameters: percent retention and percent expansion (Figure 2.2). Percent retention measures the percentage of genes duplicated by polyploidy that have survived in duplicate, and percent expansion measures the relative contributions of polyploid vs. non-polyploid [NP]) duplications to current gene family size (see Methods for details). Figure 2.2 illustrates these calculations using the *RbcS* gene family in *Glycine*.

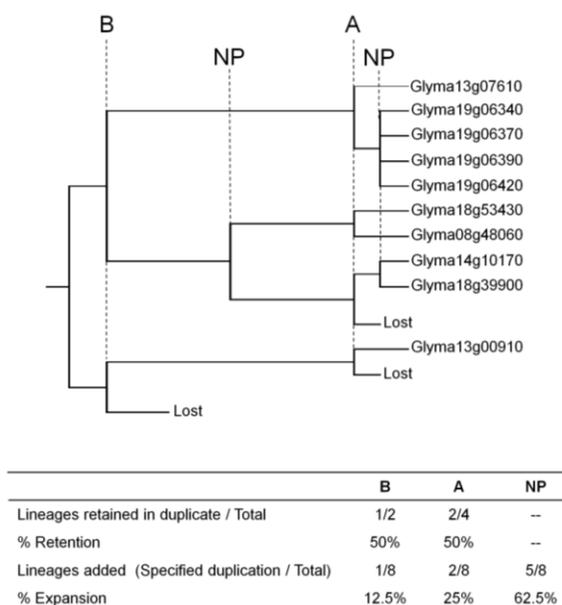


Figure 2.2. An example of percent retention and percent expansion calculations. The *RbcS* protein is encoded by ten genes in *Glycine*, dispersed on five different chromosomes (8, 13, 14, 18, and 19; Schmutz et al. 2010). Gene pairs identified as homoeologues reside in or near syntenic blocks and have the expected number of synonymous substitutions per synonymous site ( $K_s$ ) for that duplication (see Methods for details). For example, Glyma13g07610 and the tandem duplicates on chromosome 19 reside in a syntenic block spanning 50 genes on chromosome 13 and 77 genes on chromosome 19, with a mean  $K_s = 0.18$ . All of the 10 *RbcS* gene family members descended from one of two pre-B ancestors, such that the gene family has expanded by eight gene lineages. One of the two gene lineages added by the B duplication was subsequently lost, whereas two of four gene lineages added by A were subsequently lost. In addition to the B and A duplications, one non-polyploid (NP) duplication occurred between B and A, and four occurred between A and the present.

*Homoeologue retention in Glycine*: Figure 2.3 summarizes retention of polyploid duplicates by photosynthetic gene family and species. Across all photosynthetic gene families in *Glycine*, 78.7% (70/89) of pre-A gene lineages have retained duplicates from the A polyploidy event, and 84.9% (152/179) of photosynthetic genes present today have a homoeologue from the A duplication.

Funct. Group	Protein	Glycine		Medicago	Arabidopsis		
		B	A	B	$\beta$	$\alpha$	
CC	RbcS	1/2	2/4	0/1	0/1	1/1	
	PGK	0/2	2/2	0/2	0/2	1/2	
	GAPDH	0/5	6/6	0/5	0/4	2/4	
	TPI	1/2	3/3	0/2	0/2	0/2	
	FBA	2/5	5/7	0/3	1/4	2/5	
	FBPase	1/4	4/6	0/3	0/3	0/3	
	TKL	0/2	1/4	0/2	0/1	1/1	
	SBPase	0/1	1/1	0/1	0/1	0/1	
	RPE	0/2	2/2	0/2	0/2	0/2	
	PRP	0/6	2/7	0/3	1/3	0/4	
	PRK	0/1	1/1	0/3	0/1	0/1	
PSII	PsbO	1/1	2/2	0/2	1/1	1/2	
	PsbP	1/1	2/2	1/1	0/1	1/1	
	PsbQ	1/1	1/2	0/1	0/1	1/1	
	PsbR	0/1	1/1	0/1	0/1	0/1	
	PsbS	1/1	1/2	0/1	0/1	0/1	
	PsbTn	0/1	2/2	1/1	0/1	1/1	
	PsbW	1/1	2/2	1/1	0/1	0/1	
	PsbX	0/2	2/3	2/2	0/1	0/1	
	PsbY	1/1	2/2	0/2	0/1	0/1	
	PsaD	0/1	1/1	0/1	0/1	1/1	
PSI	PsaE	0/1	1/1	1/1	0/1	1/1	
	PsaF	0/1	0/1	0/1	0/1	0/1	
	PsaG	0/1	1/1	0/1	0/1	0/1	
	PsaH	1/1	2/2	0/1	0/1	1/1	
	PsaK	0/1	1/1	0/1	0/1	0/1	
	PsaL	0/1	1/1	0/1	0/1	0/1	
	PsaN	0/1	1/1	0/1	0/1	0/1	
	PsaO	0/1	1/1	0/1	0/1	0/1	
	LHC	LhcA1	0/1	1/1	0/1	0/1	0/1
		LhcA2	1/1	2/2	0/1	0/1	0/1
LhcA3		0/1	1/1	0/1	0/1	0/1	
LhcA4		0/1	1/1	0/1	0/1	0/1	
LhcA5		0/1	1/1	0/1	0/1	0/1	
LhcA6		0/1	1/1	0/1	0/1	0/1	
LhcB1		0/3	2/3	1/2	0/1	1/1	
LhcB2		0/1	1/1	0/1	0/1	0/2	
LhcB3		0/1	1/1	0/1	0/1	0/1	
LhcB4		0/2	2/2	0/2	0/2	1/2	
LhcB5		1/1	2/2	0/1	0/2	0/2	
LhcB6	1/1	2/2	0/1	0/1	0/1		
TOTAL		15/66	70/89	7/62	3/56	16/60	

100%  
67-99%  
34-66%  
1-33%  
0%

Figure 2.3. Percent retention of homoeologues, given by gene family and species, for the Calvin cycle (CC), photosystems II and I (PSII, PSI), and the light harvesting complex (LHC). Shading indicates percent retention, and values indicate the number of gene lineages retained in duplicate over the number of gene lineages initially duplicated by the specified polyploidy event.

In contrast, across the whole genome between 43.4% and 67.3% of genes have retained homoeologues from the A duplication (Schmutz et al. 2010). The upper

estimate for genome-wide retention (67.3%) includes all genes with paralogues of any age (Schmutz et al. 2010), and is, therefore, an overestimate for the A duplication (some of these paralogues resulted from the B or even older polyploidy events, including an ancient hexaploidy [Tang et al. 2008], or from NP duplications). However, even compared to this upper estimate, photosynthetic gene families exhibit a significantly higher rate of retention from the A duplication ( $\chi^2_1 = 25.083$ ,  $p = 5.5 \times 10^{-07}$ ) than the genome-wide average (Figure 2.4).

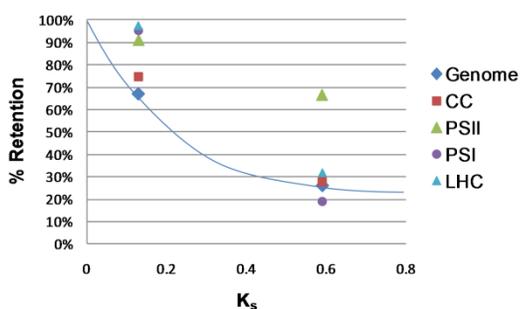


Figure 2.4. Observed retention rates for photosynthetic genes following polyploidy in *Glycine*, compared to syntenic-based genome-wide retention rates from Schmutz et al. 2010.

Though retention is high for photosynthetic gene families overall, retention rates following the A duplication differ significantly among the different functional groups (Figure 2.4). For CC genes, retention of duplicates from the A polyploidy event (74.7%) is comparable to the upper estimate (67.3%) for the genome wide average ( $\chi^2_1 = 2.149$ ,  $p = 0.143$ ). In contrast, within the three thylakoid membrane-associated protein complexes (PSII, PSI and LHC), duplicate retention is much higher (90.9%, 95.2% and 97.4%, respectively). Combined, retention for the three thylakoid-associated functional groups (94.7%) is significantly higher than for the CC ( $\chi^2_1 = 13.759$ ,  $p = 2.0 \times 10^{-4}$ ) or for the upper estimate for the genome-wide average ( $\chi^2_1 = 30.978$ ,  $p = 3.0 \times 10^{-8}$ ).

Despite significantly higher retention of A homoeologues than the genome-wide average, photosynthetic gene families overall have fractionated at a rate comparable to or below the genome-wide level following the B duplication (Figure 2.4). For all photosynthetic gene families, 22.7% (15/66) of pre-B gene lineages have retained both duplicates from the B polyploidy event, and 34.6% (62/179) of photosynthetic genes present today retain homoeologues from the B duplication. Based on internal synteny analysis, 25.9% of present-day genes genome-wide retain duplicates from the B event (Schmutz et al. 2010).

As with the A duplication, retention rates vary considerably by photosynthetic functional group following the B duplication. CC (27.6%, 24/87), PSI (19.0%, 4/21) and LHC (31.6%, 12/38) all exhibit equivalent retention rates to each other ( $\chi^2_2 = 1.072$ ,  $p = 0.5851$ ), and to the synteny-based estimate (25.9%) for the genome wide average. In contrast, duplicate retention in PSII is significantly higher (66.7%, 22/33) than in the other functional groups ( $\chi^2_3 = 19.275$ ,  $p = 2.4 \times 10^{-4}$ ) (Figure 2.4).

*Gene family expansion in Glycine*: Figure 2.5 shows the contributions of polyploidy and NP duplications to gene family expansion for each photosynthetic gene family. In *Glycine*, the fraction of gene families retaining NP duplicates differs significantly among the four functional groups ( $p = 0.02$ ; Fisher's exact test). NP duplications have made a larger contribution to the expansion of CC gene families than to the expansion of LHC, PSII or PSI (Table 2.2 and Figure 2.6). Seven of 11 CC gene families (64%) have expanded via NP mechanisms, compared to only 2/9 (22%), 1/9 (11%) and 1/12 (8%) for PSII, PSI and LHC, respectively (Figure 2.5). On average, NP duplications have contributed 27.2% of total gene family expansion in the CC, compared to 7.4%, 8.3% and 5.0% for PSII, PSI and LHC, respectively. Thus gene family expansion is strongly biased towards polyploid duplication in the

thylakoid-associated complexes, and more balanced between polyploid and non-polyploid duplications for the enzymes of the CC (Table 2.2 and Figure 2.6).

Funct. Group	Protein	Glycine			Medicago		Arabidopsis			
		B	A	NP	B	NP	$\beta$	$\alpha$	NP	
CC	RbcS	1/8	2/8	5/8	0/5	5/5	0/3	1/3	2/3	
	PGK	0/2	2/2	0/2	--	--	0/1	1/1	0/1	
	GAPDH	0/8	6/8	2/8	--	--	0/3	2/3	1/3	
	TPI	1/4	3/4	0/4	--	--	--	--	--	
	FBA	2/9	5/9	2/9	0/1	1/1	1/4	2/4	1/4	
	FBPase	1/6	4/6	1/6	--	--	--	--	--	
	TKL	0/10	1/10	9/10	--	--	0/1	1/1	0/1	
	SBPase	0/2	1/2	1/2	--	--	--	--	--	
	RPE	0/2	2/2	0/2	0/1	1/1	0/1	0/1	1/1	
	PR1	0/3	2/3	1/3	--	--	1/1	0/1	0/1	
	PRK	0/1	1/1	0/1	--	--	--	--	--	
PSII	PsbO	1/3	2/3	0/3	--	--	1/2	1/2	0/2	
	PsbP	1/3	2/3	0/3	1/1	0/1	0/1	1/1	0/1	
	PsbQ	1/2	1/2	0/2	--	--	0/1	1/1	0/1	
	PsbR	0/1	1/1	0/1	--	--	--	--	--	
	PsbS	1/2	1/2	0/2	--	--	--	--	--	
	PsbTn	0/3	2/3	1/3	1/2	1/2	0/1	1/1	0/1	
	PsbW	1/3	2/3	0/3	1/1	0/1	--	--	--	
	PsbX	0/3	2/3	1/3	2/2	0/2	--	--	--	
	PsbY	1/3	2/3	0/3	0/1	1/1	--	--	--	
	PsaD	0/1	1/1	0/1	--	--	0/1	1/1	0/1	
PSI	PsaE	0/3	1/3	2/3	1/1	0/1	0/1	1/1	0/1	
	PsaF	--	--	--	--	--	--	--	--	
	PsaG	0/1	1/1	0/1	--	--	--	--	--	
	PsaH	1/3	2/3	0/3	--	--	0/1	1/1	0/1	
	PsaK	0/1	1/1	0/1	--	--	--	--	--	
	PsaL	0/1	1/1	0/1	0/1	1/1	--	--	--	
	PsaN	0/1	1/1	0/1	--	--	--	--	--	
	PsaO	0/1	1/1	0/1	--	--	--	--	--	
	LHC	LhcA1	0/1	1/1	0/1	--	--	--	--	--
		LhcA2	1/3	2/3	0/3	--	--	0/1	0/1	1/1
LhcA3		0/1	1/1	0/1	--	--	--	--	--	
LhcA4		0/1	1/1	0/1	--	--	--	--	--	
LhcA5		0/1	1/1	0/1	--	--	--	--	--	
LhcA6		0/1	1/1	0/1	--	--	--	--	--	
LhcB1		0/5	2/5	3/5	1/4	3/4	0/4	1/4	3/4	
LhcB2		0/1	1/1	0/1	--	--	0/2	0/2	2/2	
LhcB3		0/1	1/1	0/1	--	--	--	--	--	
LhcB4		0/2	2/2	0/2	--	--	0/1	1/1	0/1	
LhcB5		1/3	2/3	0/3	--	--	--	--	--	
LhcB6		1/3	2/3	0/3	--	--	--	--	--	
TOTAL		15/113	70/113	28/113	7/20	13/20	3/30	16/30	11/30	

Figure 2.5. Percent expansion of gene families by species for the CC, PSII and PSI, and the LHC. Shading indicates percent expansion, and values indicate number of gene lineages added by the designated duplication over the total number of gene lineages added, starting with the oldest polyploidy event.

The CC has also retained more duplicates that predate the B polyploidy event than have the thylakoid-associated complexes (Figure 2.6A). We utilized Phytozome gene clusters (<http://www.phytozome.net>) to test if any of these pre-B duplications were part of the ancient hexaploidy shared by all eudicots (Tang et al., 2008).

Phytozome gene clusters were produced using a combination of BLASTP and synteny

searches in order to reconstruct ancestral gene sets for key phylogenetic nodes, including “Rosid (pre-hexaploidy)” and “Rosid (post-hexaploidy)”. *Glycine* genes that cluster at the “Rosid (pre-hexaploidy)” node, but not at more recent nodes (such as “Rosid (post-hexaploidy),” “Eurosid I” or “Legume”) were likely derived from this paleohexaploidy. None of the 21 pre-B duplications in the CC fit these criteria (15 of the 21 clustered only at older nodes, and 6 clustered at the more recent “Eurosid I” node). Additionally, the pre-B duplication in the *PGK* gene family was a tandem duplication (that was subsequently duplicated again by the A polyploidy event to yield two tandem pairs, Glyma08g17600 / Glyma08g17610 and Glyma15g41540 / Glyma15g41550), providing additional evidence against WGD in this case.

Table 2.2. Average percent retention and percent expansion following polyploidy events by functional groupings of photosynthetic gene families. Retention and expansion values were obtained by averaging the values from each gene family within a functional group.

Species	% Retention / % Expansion				
	CC (n*=11)	PSII (n=9)	PSI (n=9)	LHC (n=12)	Combined (n=41)
<i>Glycine</i> – B	15.0 / 6.9	66.7 / 25.9	11.1 / 4.2	25.0 / 8.3	28.4 / 11.0
<i>Glycine</i> – A	76.5 / 65.8	85.2 / 66.7	88.9 / 87.5	97.2 / 86.7	87.9 / 76.7
<i>Glycine</i> – NP	-- / 27.2	-- / 7.4	-- / 8.3	-- / 5.0	-- / 12.3
<i>Medicago</i> – B	0.0 / 0.0	44.4 / 70.0	11.1 / 50.0	4.2 / 25.0	13.4 / 47.3
<i>Medicago</i> – NP	-- / 100.0	-- / 30.0	-- / 50.0	-- / 75.0	-- / 52.7
<i>Arabidopsis</i> – $\beta$	5.3 / 17.9	11.1 / 12.5	0.0 / 0.0	0.0 / 0.0	6.3 / 15.3
<i>Arabidopsis</i> – $\alpha$	30.9 / 50.0	38.9 / 87.5	33.3 / 100.0	12.5 / 31.3	25.4 / 57.0
<i>Arabidopsis</i> - NP**	-- / 32.1	-- / 0.0	-- / 0.0	-- / 68.7	-- / 27.8

\*n = number of nuclear gene families associated with each functional group

\*\*NP: non-polyploid gene duplications (post- $\beta$ /B only)

In contrast to the 21 pre-B duplications observed in the CC, only four pre-B duplications were detected across the three thylakoid-associated functional groups.

Two of these (one in the PSII gene family, *PsbX*, and one in the LHC gene family, *LhcB4*) cluster at nodes more recent than “Rosid (pre-hexaploidy),” and the other two (both in the *LhcB1* gene family) cluster at older nodes. Thus, we find no evidence that any of the pre-B duplications in photosynthetic gene families were the result of the paleohexaploidy, and were most likely NP duplications. The greater number of retained pre-B duplicates in the CC compared to the thylakoid-associated complexes is, therefore, consistent with the greater number of post-B NP duplications in the CC.

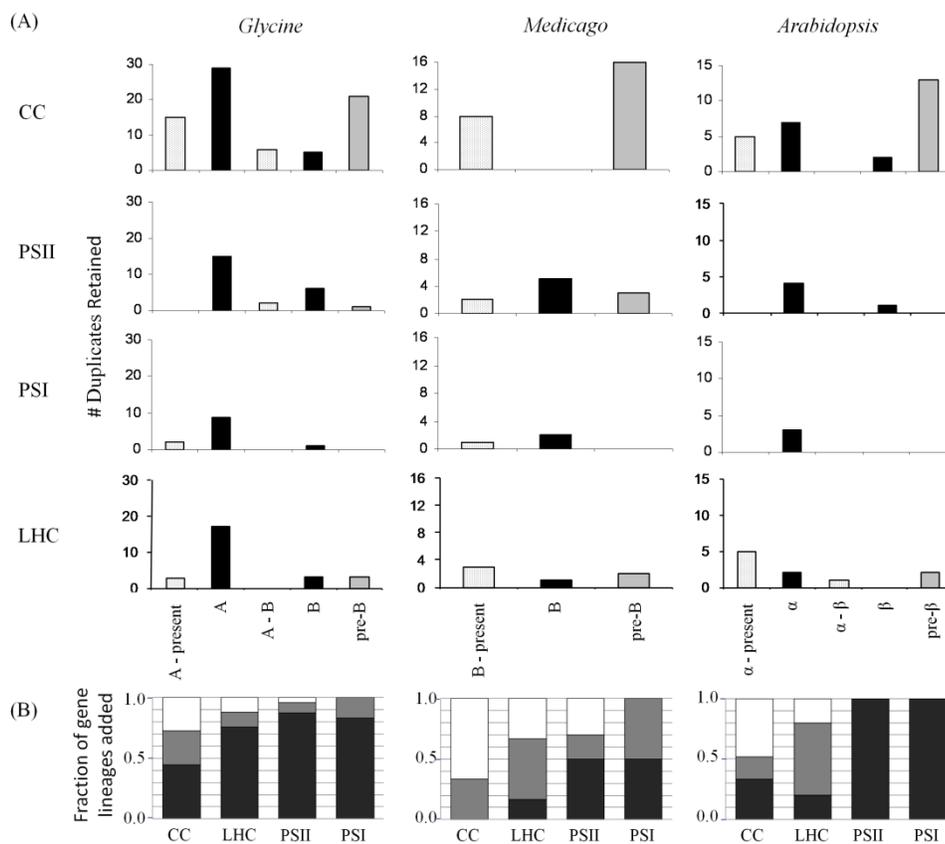


Figure 2.6. (A) Gene duplicates retained by duplication category and functional group in *Glycine*, *Medicago* and *Arabidopsis*. P duplicates are shown in black, NP duplicates are shown in stippled white, and ancient duplicates are shown in gray. (B) Fraction of gene lineages added by duplication type (P, NP, Pre-B/β) for the four photosynthetic functional groups in *Glycine*, *Medicago* and *Arabidopsis*. P duplicates are shown in black, NP duplicates are shown in gray, and ancient duplicates are shown in white.

In *Glycine*, therefore, PSII exhibits high levels of retention following polyploid duplication, and minimal contribution from NP duplication to gene family expansion. In contrast, the CC exhibits the opposite pattern, with comparatively low levels of polyploid duplicate retention, and a greater contribution from NP duplication, including a large number of pre-B NP duplications, to gene family expansion. PSI and the LHC exhibit intermediate patterns, with high retention of duplicates from the A WGD, comparable to PSII, but low retention of duplicates from the B WGD, comparable to the CC. As with PSII, NP duplications have made little contribution to gene family expansion in PSI and the LHC.

*Patterns of retention and expansion across species:* The B polyploidy event took place in the common ancestor of *Glycine* and *Medicago*, shortly before the two lineages diverged (Pfeil et al. 2005; Figure 2.1A). We examined retention and expansion of photosynthetic gene families in *Medicago* to see if similar patterns developed following the same WGD in an independently evolving lineage. As with *Glycine*, PSII exhibits the highest average rate of retention of B duplicates, with the other three functional groups exhibiting notably lower retention rates (Table 2.2). Also as with *Glycine*, CC gene families exhibit the highest contributions of non-polyploid duplications to gene family expansion in *Medicago*, as well as the greatest number of pre-B duplications (Table 2.2 and Figure 2.6). Thus, consistent with the patterns observed in *Glycine*, gene family expansion is biased towards polyploid duplication in the two photosystems (and, to a lesser extent, the LHC), and more balanced between polyploid and non-polyploid duplications for the CC (Tables 2.2 and Figure 2.6) in *Medicago*.

It should be noted that, overall, we observed lower duplicate retention from B in *Medicago* than in *Glycine* (Figure 2.3), which may be due to incomplete sampling

of homoeologues in *Medicago*. Whereas our retention estimates for *Glycine* are based on conserved synteny within a relatively complete genome sequence, due to the incomplete nature of the *Medicago* genome sequence (the *Medicago* Genome Sequence Consortium estimates that 70% of gene space has been sequenced in release 3.0; N. Young, pers. comm.) our inferences about retention for this species are based primarily on  $K_s$  and gene tree topology estimates from tentative consensus (TC) sequences. In *Glycine*, retention estimates derived from TCs were identical to estimates from synteny for 35 of 41 gene families, but TC-based estimates were lower than synteny-based estimates for five of the six families that differed. Thus our retention estimates for *Medicago* are likely underestimates. However, these limitations should affect all four functional groups equally, and not affect differences in overall pattern among functional groups.

The lineage leading to *Arabidopsis* experienced two polyploidy events, designated  $\beta$  and  $\alpha$  (Bowers et al. 2003; Figure 2.1A) subsequent to the ancient hexaploidy at the base of the eudicots, and after divergence from the lineage (Eurosid I) that gave rise to legumes. We therefore examined duplicate retention and expansion of photosynthetic genes in *Arabidopsis* in order to see if the patterns observed in the two legume species also emerged following these completely independent (and older) duplications (Figure 2.1).

Across all photosynthetic gene families in *Arabidopsis*, 26.7% (16/60) of pre- $\alpha$  gene lineages have retained duplicates from the  $\alpha$  polyploidy event, and 43.0% (37/86) of photosynthetic genes present today have a homoeologue from the  $\alpha$  duplication. After collapsing recent tandem duplicates into a single locus, as per Thomas et al. (2006), 40.5% (32/79) of photosynthetic genes present today retain homoeologues from the  $\alpha$  duplication. In contrast, across the whole genome, 28.5% (6,329/22,209) of genes retain  $\alpha$  homoeologues (Thomas et al. 2006). Thus, as in

*Glycine*, photosynthetic genes in *Arabidopsis* have significantly higher retention of  $\alpha$  duplicates than the genome-wide average ( $\chi^2_1 = 5.566$ ,  $p = 0.018$ ).

Also consistent with *Glycine*, retention rates following the  $\alpha$  duplication differ significantly among the different functional groups in *Arabidopsis*. Again, the CC has a retention rate (40.0%) comparable to the genome-wide average ( $\chi^2_1 = 2.268$ ,  $p = 0.132$ ), whereas PSII (57.1%) is significantly higher ( $\chi^2_1 = 4.314$ ,  $p = 0.038$ ; Yate's correction). PSI (50%) also exhibits higher retention than the genome-wide average, though due to the small numbers of genes involved, the difference is not statistically significant ( $\chi^2_1 = 1.768$ ,  $p = 0.184$ ). In contrast to *Glycine*, the LHC has low retention of  $\alpha$  homoeologues in *Arabidopsis* (22.2%), comparable to the genome-wide average ( $\chi^2_1 = 0.348$ ,  $p = 0.555$ ).

Again consistent with the pattern observed in *Glycine*, despite higher retention of duplicates following the  $\alpha$  polyploidy event, photosynthetic genes in *Arabidopsis* exhibit retention rates following the  $\beta$  duplication that are comparable to or lower than the genome-wide average. Across all photosynthetic gene families, only 10.1% (8/79) retain  $\beta$  homoeologues. Based on a model of decay rate following the  $\beta$  duplication (Maere et al. 2005), genome-wide retention is approximately 13.8% for this event. Using a different approach, Bowers et al. (2003) found that 21.4% of genes (2,874/13,449) retain homoeologues within duplicated blocks resulting from the  $\beta$  duplication. Compared to this estimate of genome-wide retention, PS genes have significantly lower retention rates ( $\chi^2_1 = 5.922$ ,  $p = 0.015$ ).

Consistent with the two legume lineages, PSII exhibits the highest retention rate from the polyploidy events in *Arabidopsis* (Table 2.2). Also consistent with the legume lineages, the CC has the highest fraction of gene families (4/11) that have expanded via NP duplication, and higher percent expansion via NP duplications than either photosystem (Table 2.2 and Figure 2.5). In addition, as in the legume species,

CC gene families have retained more pre- $\beta$  duplicates than any of the thylakoid-associated functional groups in *Arabidopsis* (Figure 2.6). Tang et al. (2008) generated gene clusters, including *Arabidopsis* genes, representing ancestral genes that were duplicated by the ancient hexaploidy event. We checked to see if any pre- $\beta$  duplications collapse into these gene clusters, indicating that they were the result of the hexaploidy event. Of the 13 pre- $\beta$  duplications within the CC, only one (in *PGK*) could be assigned to the hexaploidy. Thus, the majority of these pre- $\beta$  duplications were likely also NP duplications. Of the two pre- $\beta$  duplications in LHC gene families, one (*LhcB4*) was assigned to the hexaploidy. Neither photosystem retains pre- $\beta$  duplications.

Thus, consistent patterns emerge across three species and two independent sets of whole genome duplications. First, photosynthetic genes overall have higher retention of duplicates from the most recent polyploidy events than the genome-wide average, though CC genes exhibit retention rates comparable to the genome-wide average. Second, following older polyploidy events, photosynthetic genes exhibit fractionation comparable to or greater than the genome-wide average. Third, of the photosynthetic functional groups, PSII exhibits the highest retention of polyploid duplicates in the long term (Table 2.2). Fourth, polyploid duplications have contributed more to gene family expansion in both photosystems than in the CC (Figure 2.6B). Fifth, the CC exhibits the highest level of gene family expansion via NP duplication (Figure 2.6), including very old NP duplications that predate the B/ $\beta$  polyploidy events.

*Patterns of duplicate retention and expansion at the gene family level:* We next asked if patterns that emerge at the level of photosynthetic functional groups are due to consistent patterns of duplicate retention and loss at the level of individual gene families. Because the *Glycine* genome has experienced two successive WGD events,

we looked to see if patterns of retention and loss are consistent across the two duplications. However, because of the very high level of duplicate retention from the A polyploidy event within photosynthetic gene families (40 of 41 gene families have retained at least one homoeologue and 31/41 have retained 100% of homoeologues from this event; Figure 2.3), any fractionation patterns that might develop over longer evolutionary timeframes would not yet be apparent. Thus, such an analysis is necessarily inconclusive. To partially circumvent this issue, we focused on those gene families that have undergone partial or complete fractionation from the A duplication. 10 gene families have lost one or more homoeologues from A (Figure 2.3). Of these 10, eight (80%) have also lost one or more homoeologues from B. In comparison, 23 of 31 (74%) photosynthetic gene families that have retained all homoeologues following the A duplication have lost one or more homoeologues from the B duplication. Thus, gene families that have lost duplicates from the A polyploidy event were no more likely to have also lost duplicates from the B event than gene families that have retained all A duplicates ( $\chi^2_1 = 0.003$ ,  $p = 0.96$ ; Yates' correction for continuity).

We then looked to see if patterns emerge at the level of individual gene families when also considering the shared duplication (B) in *Medicago*, and the two independent duplications ( $\beta$  and  $\alpha$ ) in *Arabidopsis*. When looking across species, there is a striking absence of pattern in terms of homoeologue retention, with only four of 41 gene families retaining duplicates from polyploidy in all three species, and only one of 41 families (*PsaF*) eliminating all polyploid duplicates in all three species.

As mentioned above, long-term patterns of retention and loss could be obscured by under-sampling of homoeologues in *Medicago* and by the high level of retention following the A polyploidy event in *Glycine*. If, however, we restrict our comparisons to only consider the two *Arabidopsis* duplications and the B duplication

in *Glycine*, we still find a striking absence of pattern in homoeologue retention. Of the 23 gene families that have retained homoeologues in either species, only six have retained duplicates in both species (Figure 2.3). Limiting our comparisons to the two WGD's in *Arabidopsis*, of the 15 photosynthetic gene families that have retained duplicates from at least one polyploidy event, only two have retained duplicates from both  $\alpha$  and  $\beta$  (Figure 2.3). Comparing percent retention values across the 41 gene families, we observe negligible correlation ( $r \leq 0.23$ ) when comparing the B polyploidy event in *Glycine* to either polyploidy event in *Arabidopsis*, or when comparing the two nested duplications within *Arabidopsis*.

Similarly, few consistent patterns of expansion were apparent across species, or across nested duplication events within species (Figure 2.5). Looking at all three species and four polyploidy events, of the seven gene families that have expanded in size in all three species, only four have expanded by polyploidy in all three, and only one (*PsbP*) has expanded exclusively by polyploidy. Restricting the comparison to *Arabidopsis* and the B duplication in *Glycine*, of 13 gene families that have expanded in size in both species, only six have expanded by polyploidy in both, and only four have expanded exclusively by polyploidy.

Conversely, only three gene families have expanded by NP duplications in all three species (*LhcB1*, *RbcS* and *FBA*), and none of these have expanded exclusively by small-scale mechanisms. Again restricting the comparison to *Arabidopsis* and the B event in *Glycine*, only four gene families have expanded by NP duplication (*LhcB1*, *RbcS*, *FBA*, and *GAPDH*) and none have expanded exclusively via NP mechanisms in both species.

*Assessment of functional divergence:* Consistent differences in retention and expansion at the level of the four photosynthetic functional groups suggest that different evolutionary forces are acting upon duplicates within each group. Because a

common explanation for duplicate retention is functional differentiation, we looked for evidence of positive selection and/or expression divergence between duplicated photosynthetic genes.

Global  $\omega$  ( $K_a/K_s$ ) was measured for gene pairs resulting from each duplication mechanism. In all cases, all photosynthetic gene family members in all four classes appear to be under purifying selection in all three taxa ( $\omega < 1$ ) (Table 2.3). We then looked for local signatures of positive selection within sliding windows of both sequence and spatial domains (windows of either 30 adjacent codons in the primary sequence or windows of amino acid residues contained within 10 Å spheres in the folded protein) for duplicates from the most recent (A/ $\alpha$ ) polyploidy events in *Glycine* and *Arabidopsis*. Using these more sensitive approaches, we found evidence for positive selection ( $\omega > 1.2$ ) within local domains for several gene families. In *Glycine*, the majority of A homoeologue pairs of CC genes (25 of 28) show evidence of positive selection, including duplicates from every gene family except *PRK*, whereas fewer than half of photosystem or LHC homoeologues exhibit signatures of positive selection (Table 2.3 and Figure 2.7A).

Table 2.3. Sliding window estimates of selection ( $\omega$ ) by functional group for the most recent polyploidy events ( $\alpha$  and A) in *Arabidopsis* and *Glycine*.

Species	Funct. Group	Global $\omega$ (Avg.)	Sliding Window $\omega$		
			Fraction of homoeologue pairs exhibiting pos. selection		
			1D*	3D*	Combined**
<i>Glycine</i>	CC	0.12	0.36 (10/28)	0.92 (24/26)	0.89 (25/28)
	PSII	0.18	0.13 (2/15)	0.86 (6/7)	0.47 (7/15)
	PSI	0.20	0.20 (2/10)	0.38 (3/8)	0.40 (4/10)
	LHC	0.10	0.20 (3/15)	0.71 (5/7)	0.47 (7/15)
<i>Arabidopsis</i>	CC	0.09	0.00 (0/7)	0.43 (3/7)	0.43 (3/7)
	PSII	0.13	0.00 (0/4)	0.33 (1/3)	0.33 (1/3)
	PSI	0.14	0.00 (0/3)	0.33 (1/3)	0.33 (1/3)
	LHC	0.10	0.00 (0/2)	0.00 (0/1)	0.00 (0/2)

\*1D = 1-dimensional sliding window (30aa); 3D: 3-dimensional window (10 Å)

\*\*Fraction of pairs exhibiting positive selection in 1D and/or 3D analysis

In *Arabidopsis*, three of seven  $\alpha$  homoeologue pairs of CC genes exhibit signatures of positive selection (*RbcS*, *PGK* and *GAPDH*). In each of the two photosystems, positive selection was detected for one of three homoeologue pairs (*PsaH* from PSI and *PsbQ* from PSII). Only two  $\alpha$  homoeologue pairs remain for LHC genes, and no evidence of positive selection was detected for either (Figure 2.7A).

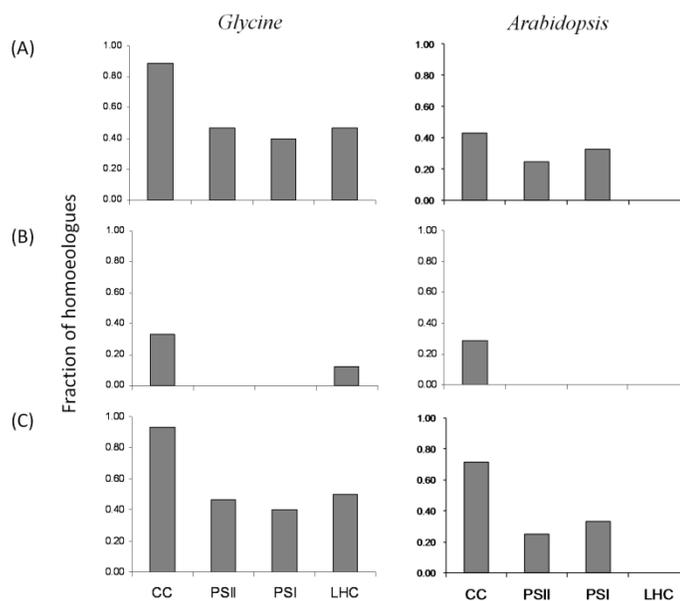


Figure 2.7. Fractions of homoeologue pairs from the most recent polyploidy event ( $\alpha$  or A) exhibiting evidence for functional divergence. (A) Fractions of homoeologous duplicates exhibiting signatures of positive selection (sliding window  $\omega$ ), (B) divergence in expression profiles ( $r < 0.5$ ), (C) or some combination of the two, separated by species (*Arabidopsis* or *Glycine*) and functional group (CC = Calvin cycle; PSII = photosystem II; PSI = photosystem I; LHC = light harvesting complex).

We explored the expression profiles of duplicated photosynthetic genes using data from several RNA-Seq experiments in *Glycine* (Bolon et al. 2010, Libault et al. 2010). CC duplicates exhibit lower average correlation coefficients than photosystem or LHC duplicates, regardless of duplication mechanism (Table 2.4). For example, all duplicate pairs from the A polyploidy event maintain highly correlated expression

profiles in PSI ( $r \geq 0.85$ ) and PSII ( $r \geq 0.94$ ). LHC gene pairs from the A duplication exhibit a somewhat greater diversity in expression profiles, with 15 of 17 A duplicates highly correlated ( $r \geq 0.80$ ), but two pairs (from *LhcB1* and *LhcB6*) showing evidence of expression divergence ( $r < 0.3$ ). In contrast, for 28 CC gene pairs from the A duplication for which both copies are expressed, eight exhibit divergent expression profiles ( $r < 0.55$ ), including negative values for two pairs (from *TPI* and *PRI*) (Figure 2.7B).

We also explored the expression profiles of duplicated photosynthetic genes in *Arabidopsis* using public microarray data. Similar to *Glycine*, duplicate genes from both photosystems exhibit highly correlated expression profiles ( $r > 0.9$ ) regardless of duplication type (Table 2.4). The only exceptions are a pair of  $\alpha$  homoeologues for *PsbP* ( $r = 0.55$ ) and a pair of  $\beta$  homoeologues for *PsbO* ( $r = 0.08$ ). In both cases, one of the two copies is effectively silent in all tissues and conditions examined ([www.genevestigator.org](http://www.genevestigator.org); data not shown). Also similar to *Glycine*, LHC duplicates exhibit a slightly greater diversity of expression profiles. Pre- $\beta$  duplicates of *LhcB4* and *LhcB5* genes exhibit lower levels of co-expression ( $r = 0.68$  and  $0.52$  respectively), but the three more recent duplicates for which expression profiles can be discriminated are highly co-expressed ( $r_{\text{avg}} = 0.93$ ,  $r_{\text{min}} = 0.90$ ). The pre- $\beta$  *LhcB5* duplication involves a gene (AT1G76570) that appears to encode a structurally distinct LHC protein, with four transmembrane (TMM) domains instead of the canonical three (making it more like the four TMM-protein, *PsbS*) (Klimmek et al., 2006). This gene and one of the genes involved in the pre- $\beta$  *LhcB4* duplication exhibit expression profiles more like *PsbS*. Due to these structural and expression differences, Klimmek et al. (2006) proposed to reclassify these genes into distinct families (*LhcB7* and *LhcB8*).

In contrast to the uniformly high level of co-expression within the two photosystems and LHC, *Arabidopsis* CC duplicates, as in *Glycine*, exhibit considerable variation in degree of co-expression, with Pearson correlation coefficients ranging from -0.67 to 0.98. On average, expression profiles are less correlated for CC duplicates than for duplicates of photosystem or LHC genes, regardless of the mechanism of duplication (Table 2.4).

Table 2.4. Degree of expression correlation between duplicate gene pairs by duplication type and functional group in *Arabidopsis* and *Glycine*.

Species	Funct. Group	Average (Min. / Max.) expression correlation (r)				
		Pre-B	B	B - A	A	A - present
<i>Glycine</i>	CC	0.00 (-0.41/0.28)	0.57 (0.27/0.83)	0.89 (0.85/0.92)	0.73 (0.20/1.00)	0.68 (0.04/0.92)
	PSII	0.87	0.68 (-0.13/0.97)	0.95 (0.94/0.96)	0.98 (0.94/1.00)	--
	PSI	--	0.78	--	0.96 (0.85/1.00)	0.95
	LHC	0.71 (0.63/0.79)	0.76 (0.55/0.92)	--	0.88 (0.58/0.99)	0.34
		Pre- $\beta$	$\beta$	$\beta$ - $\alpha$	$\alpha$	$\alpha$ - present
<i>Arabidopsis</i>	CC	0.15 (-0.62/0.97)	-0.05 (-0.05/0.06)	--	0.44 (-0.67/0.98)	0.46 (0.27/0.74)
	PSII	--	--	--	0.94(0.91/0.96)	--
	PSI	--	--	--	0.97 (0.95/0.98)	--
	LHC	0.60 (0.52/0.68)	--	0.90	0.94 (0.94/0.95)	?

--: no duplicates retained; ?: all duplicate pairs hybridize to the same Affymetrix probes so that degree of co-expression could not be determined using Affymetrix array data. No min. or max. values are given for duplication categories with only one retained pair.

Extending the analysis of co-expression beyond duplicates within gene families, all genes encoding subunits of PSI are highly co-regulated (for all pairwise comparisons of PSI genes,  $r_{\text{avg}} = 0.96$ , and  $r_{\text{min}} = 0.94$ ; Figure 2.8). Excluding the silent PSII homoeologues, all genes encoding subunits of PSII are also highly co-regulated ( $r_{\text{avg}} = 0.95$ ,  $r_{\text{min}} = 0.83$ ). LHC genes exhibit a greater diversity of expression profiles ( $r_{\text{avg}} = 0.85$ ,  $r_{\text{min}} = 0.51$ ), but most of this diversity results from the expression profiles of the unusual *LhcB5* (AT1G76570), and, to a lesser extent, *LhcB4*

(AT2G40100) genes. Otherwise LHC genes are co-expressed, though not to the same extent as are photosystem genes ( $r_{\text{avg}} = 0.91$ ,  $r_{\text{min}} = 0.71$ ). In contrast, CC genes show a greater variety of expression patterns (Figure 2.8). In general, CC genes cluster into two general expression profiles, but even within these two clusters there is a greater range of correlation coefficients than is found within PSI, PSII or the LHCs.

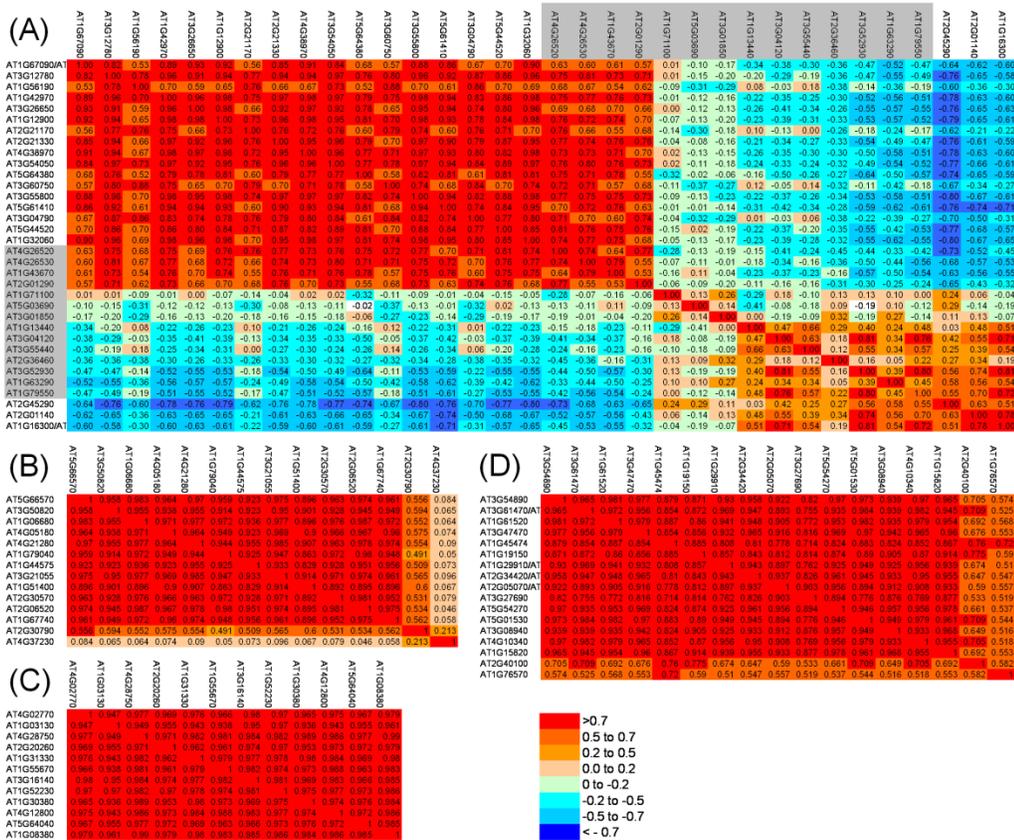


Figure 2.8. Heat maps of expression correlation coefficients ( $r$ ) within photosynthetic functional groups in *Arabidopsis*. (A) CC; (B) PSII; (C) PSI; (D) LHC. CC gene family members whose gene products localize to the cytosol are shaded in grey.

Combining data on selection and expression, greater than 70% of all  $\alpha/A$  homoeologs of CC genes exhibit some evidence for functional divergence (positive

selection, expression divergence or both), compared to  $\leq 50\%$  for PSI, PSII and LHC, in both *Glycine* and *Arabidopsis* (Figure 2.7).

## DISCUSSION

Distinct patterns of duplicate retention and gene family expansion emerge when photosynthetic gene families are considered in their functional contexts. In particular, core photosystem II (PSII) gene families exhibit relatively high levels of homoeologue retention, and both photosystems exhibit low levels of non-polyploid duplicate retention in comparison to the Calvin cycle (CC) in all three species.

This reciprocal pattern of duplicate retention suggests that different evolutionary forces are acting upon duplicates from the different photosynthetic functional groups. What might be the driving forces for these differences in duplicate retention? The pattern observed for PSII of high retention rates following genome duplications and low retention rates following single-gene duplications is the hallmark of “balanced gene drive” (Freeling and Thomas 2006). According to the “balance hypothesis,” (Papp et al. 2003) genes whose products function in multi-subunit complexes or signaling cascades will tend to be dosage-sensitive because changes in the stoichiometry of individual subunits lead to improper assembly and/or function of the complex, with deleterious consequences for the individual (Birchler et al. 2007; Birchler and Veitia 2010; Papp et al. 2003). In support of this hypothesis, Papp et al. (2003) showed in yeast that genes causing haploinsufficiency are more than twice as likely to function in complexes than genes that do not. Similarly, the class of yeast genes that are lethal when over-expressed is significantly enriched for genes whose products function in complexes.

Such dosage-sensitivity leads to distinct predictions about the retention of genes duplicated by small-scale processes vs. those duplicated by polyploidy. Small-scale duplications that affect some but not all genes encoding subunits of a protein complex will be deleterious, and should be actively eliminated from the genome by purifying selection to maintain gene balance. In contrast, whole genome duplications affect all subunits of dosage-sensitive complexes (and are, therefore, referred to as “balanced” duplications; Papp et al. 2003). Consequently, dosage-sensitive genes duplicated by polyploidy should be maintained following polyploidy, again by purifying selection for gene balance. Numerous studies have demonstrated that polyploid genomes are enriched for genes whose products function in protein-protein complexes (e.g., ribosomal proteins, proteasomal proteins and transcription factors) (Blanc and Wolfe 2004b; Freeling and Thomas 2006; Maere et al. 2005; Paterson et al. 2006; Seoighe and Gehring 2004).

The PSII complex is a large, multi-subunit protein complex with a fixed subunit stoichiometry (Minagawa & Takahashi 2004), and incompletely assembled or misassembled PSII complexes not only impair photosynthetic electron transport, but also sensitize the plant to photooxidative damage (Hwang et al. 2002, Baena-Gonzalez and Aro 2008). We suggest that these properties make PSII genes dosage sensitive. Our observation that among the four photosynthetic functional groups, PSII exhibits the highest retention rates following polyploidy, and among the lowest retention rates for non-polyploid duplications, is consistent with this hypothesis (see below for a discussion of PSI).

In contrast, non-polyploid duplications are observed more frequently in CC gene families than in either photosystem in all three species, suggesting that single-gene duplications are less likely to be deleterious in the context of the CC (as well as glycolysis and the OPPP). These observations are consistent with previous studies

demonstrating that enzymes in general tend to be dosage-insensitive (Kondrashov and Koonin 2004).

Several CC duplicates have been retained for long periods of time. Based on the  $K_s$  distributions of duplicated genes in a variety of species, the fate of the vast majority of duplicated genes is nonfunctionalization within a few million years (Lynch and Conery 2000). If these gene families are not dosage-sensitive, why have so many duplicates persisted for so long? One possibility is that there may in fact be a selective advantage to increased dosage. Two CC enzymes (*Rbcs* and *SBPase*) are thought to be rate-limiting or near-rate-limiting in carbon fixation (Harrison et al. 1996; Sun et al. 2003), and over-expression of *SBPase* increases photosynthetic rates in tobacco (Lefebvre et al., 2005; Miyagawa et al. 2001). Notably, though, *SBPase* is single-copy in *Arabidopsis* and *Medicago*, so clearly in these taxa at least there has not been strong selection for increased dosage of this gene family. The *Rbcs* gene family, in contrast, has expanded via both polyploidy and small-scale duplications in *Arabidopsis* and *Glycine*. We did not find evidence of retention of duplicates from polyploidy in *Medicago*, but its *Rbcs* gene family has expanded via recent single-gene duplications. Thus, it may be advantageous to increase gene dosage for *Rbcs*, and possibly for other CC enzymes.

Alternatively, CC duplicates may have been retained because they evolved to serve new roles in the plant. Presumably due to genetic redundancy, many duplicate gene pairs experience a period of relaxed selective constraint (Lynch and Conery 2000), which could facilitate subfunctionalization (partitioning of ancestral functions between paralogues), neofunctionalization (the acquisition of new functions by one or both paralogues), or escape from adaptive conflict (improvement of ancestral functions that were constrained when carried out by a single, ancestral gene; Des Marais and Rausher 2008). Subfunctionalized genes are retained because both copies

are required to carry out the full suite of ancestral functions. Neofunctionalized genes and genes that have undergone escape from adaptive conflict are retained if the novel or improved functions confer a selective advantage for the host.

We looked for evidence of functional differentiation in the form of positive selection and/or divergence in expression profiles. Of the four photosynthetic functional groups, we found that CC A/ $\alpha$  duplicates are the most likely to contain regions under positive selection and/or exhibit divergence in expression profiles in both *Glycine* and *Arabidopsis*. Thus, in general, it appears that photosystem genes are under strong purifying selection, and are constrained to a narrow range of correlated expression profiles, whereas CC gene families are more likely to exhibit functional divergence.

Eight of 11 CC gene families encode enzymes that function in other pathways (either glycolysis or the oxidative pentose phosphate pathway [OPPP]), and these alternative pathways may provide “functional sinks,” or avenues for sub- or neofunctionalization that facilitate retention of duplicated genes. Both the glycolytic and OPPP pathways are at least partially duplicated and spatially separated in plants, with distinct enzyme complements functioning in the plastid and cytosol in each pathway (Tobin and Bowsher 2005). Within these different compartments, the two pathways exhibit multiple levels of regulation, and the amounts and activities of the various enzymes change with tissue type and developmental stage (Tobin and Bowsher 2005). Thus, it seems plausible that greater opportunity exists for sub- and/or neofunctionalization amongst duplicates of genes encoding enzymes that function in glycolysis or OPPP in addition to the CC, compared to enzymes restricted to the CC alone. Consistent with this hypothesis, the two smallest CC gene families in *Arabidopsis*, *Glycine* and *Medicago* (*SBPase* and *PRK*) function exclusively in

photosynthetic carbon-fixation, whereas the two largest CC gene families (*GAPDH* and *FBA*) also participate in glycolysis.

In *Glycine*, A homoeologues of plastid-targeted and cytosolic CC genes exhibit comparable propensities for expression divergence (25% and 33%, respectively), and both are more likely to exhibit divergent expression patterns than are genes from either photosystem. A-homoeologues of both plastid-targeted and cytosolic CC genes are also more likely to have experienced positive selection than genes from either photosystem. So although dual-function CC gene families may have additional avenues for functional divergence than their single-function counterparts, all CC gene families appear more able to diverge functionally than the gene families of the thylakoid-associated complexes (PSI, PSII, and LHC). This, in combination with the dosage insensitivity of enzymes, could explain the relatively greater retention of non-polyploid duplicates in the CC than in the thylakoid complexes.

The fact that PSII exhibits consistently higher retention of homoeologues than the CC, despite no obvious differentiation in function between duplicates, further supports the hypothesis that these duplicated genes were simply locked in place by dosage sensitivity among the subunits of the PSII complex. This is not to say, however, that our analyses prove an absence of functional differentiation among the duplicates of PSII genes. Obviously, an overall positive correlation of expression between two genes does not preclude the possibility that the two copies have sub- or neofunctionalized at some finer scale. Indeed, careful molecular analyses of several PSII gene families have revealed differences in function. For example, using *Arabidopsis* T-DNA knockouts, Lundin et al. (2007) demonstrated that one copy of *PsbO* is more efficient than the other at supporting the oxygen-evolving capacity of PSII under photoinhibitory conditions, whereas the second copy regulates turnover of the D1 protein during the damage-repair cycle. Using transcriptional reporter gene

fusions, Sawchuk et al. (2008) showed that *LhcB2.1* (AT2G05100) is expressed at the onset of subepidermal leaf tissue development, whereas *LhcB2.3* (AT3G27690) is not expressed until late in mesophyll differentiation. It is not unlikely that other PSII gene duplications have led to functional specialization as well. Our data indicate, however, that the realm of possibilities is narrower for gene families that function strictly in photosynthesis, than for CC gene families whose products participate in multiple pathways.

If gene balance requirements explain the relatively high retention rates of homoeologues in PSII, why are duplicates retained for some PSII gene families but not others? Perhaps only a subset of the protein-protein interactions within the greater PSII complex are dosage sensitive. However, if this were the case, then we would expect to see homoeologues from the same gene families retained across nested duplications, and across all three species, yet we do not. Furthermore, in *Arabidopsis*, one  $\beta$  homoeologue from *PsbO* and one  $\alpha$  homoeologue of *PsbP* are silent, or nearly so. Obviously, these genes are not contributing to the balance of gene products (proteins).

Previous studies that supported the Balance hypothesis have demonstrated a greater propensity for “connected” genes to be retained following polyploidy (e.g., Papp et al. 2003), but this is not to say that all such genes are retained. Similarly, not all unbalanced changes in dosage involving “connected” genes are deleterious. In yeast, 37% of the genes with minimal fitness deficiency as heterozygous knockouts are involved in protein complexes (Papp et al. 2003). This highlights a key challenge associated with the Balance hypothesis – determining what precisely makes a gene dosage sensitive. Participation in a multi-protein complex alone is only weakly predictive. Recent studies at the protein level suggest that dosage sensitivity correlates with topological position within a protein complex (Veitia 2005) and with degree of

protein “under-wrapping” (the degree to which protein structural integrity is dependent on its interactive context; Liang et al. 2008), but the molecular basis for dosage sensitivity remains poorly understood.

For those genes that do indeed have dosage-related effects on fitness, dosage balance requirements are likely to be circumvented over time via other mechanisms (Aury et al. 2006, Ha et al. 2009). For example, changes in regulatory sequences, or abundance of trans-acting factors might eventually allow for balance in gene products to be achieved without maintaining balance in gene copy number. Due to multiple levels of transcriptional, post-transcriptional, translational, and post-translational control, the amount of protein product produced in many cases is likely to become at least partially uncoupled from gene dosage.

PSI exhibits similarly low levels of non-polyploid (NP) duplicate retention as PSII, consistent with dosage-sensitivity, yet also exhibits relatively low homoeologue retention rates, comparable to the CC. However, with the exception of *PsaF*, genes encoding PSI subunits exhibit 100% retention from the A polyploidy event in *Glycine* (compared to 76.5% for the CC), suggesting that there is some delay in homoeologue loss within PSI. The PSI complex therefore may be sufficiently dosage-sensitive for NP duplications to be selected against, but less sensitive than PSII (because, for example, misassembled PSI complexes do not act as sensitizers to photooxidative damage in the way that misassembled PSII complexes do). Alternatively, changes in gene dosage may be more readily adjusted for at the level of expression, allowing for a quicker decay of homoeologues despite initial dosage sensitivity. The PSII complex has more total subunits than PSI (approximately 25 vs. 15), and nearly two thirds of the PSII subunits are encoded by the plastid (14 chloroplast-encoded subunits vs. nine that are nuclear-encoded), compared to only one third for PSI (five chloroplast vs. nine nuclear). Chloroplast number per cell increases (typically doubling) with genome

doubling (Warner and Edwards 1993), thereby increasing demand for nuclear-encoded subunits of chloroplast-localized protein complexes. Coordination of nuclear- and plastid-encoded subunit stoichiometry might, therefore, represent a more significant challenge in PSII than in PSI, thereby driving longer-term retention of balanced duplicates in PSII than in PSI.

Like the CC, LHC gene families tend to exhibit lower retention of homoeologues, and greater retention of NP duplicates than the photosystems. This suggests that the LHC is not dosage sensitive. Unlike the CC, however, we find relatively little evidence for functional differentiation among LHC duplicates, even those that have been retained since before the  $\beta$ /B polyploidy events. It is possible that these gene families are dosage-sensitive, but only weakly so; or, as we speculate for PSI, that they are able to rapidly “correct” changes in gene dosage at the level of expression. Thus, selection could be too weak or too short lived to result in elevated levels of homoeologue retention, or to stringently eliminate unbalanced NP duplicates. All LHC proteins are encoded by the nucleus, so if coordination of nuclear- and plastid-encoded subunit stoichiometry prolongs gene balance constraints, LHC homoeologues would be expected to decay more rapidly than PSII homoeologues.

Alternatively, the LHC may be dosage-insensitive, and duplicate retention could be driven by functional differentiation that our analyses failed to detect. In *Arabidopsis*, most of the NP duplicates in the LHC resulted from recent duplications, and Affymetrix microarray probes do not discriminate amongst LHC paralogues. Thus, we were unable to compare the expression profiles of the recent tandem duplicates for *LhcA2*, *LhcB1* or *LhcB2*. Using transcriptional reporter fusions, Sawchuk et al. (2008) have shown differences in expression domains across tissue types and developmental stages for duplicated *Lhc* genes.

Intriguingly, the *LhcB1* gene family has experienced independent tandem duplications in all three lineages analyzed here, *LhcB2* has undergone recent tandem duplication in *Arabidopsis*, and each of the major *Lhc*'s (*LhcB1-3*) has undergone recent tandem duplications in tomato (Cannon et al. 2004). The major Lhc proteins are the most abundant proteins in the light harvesting complex, and LhcB1 and LhcB2 play important roles in balancing excitation pressure between the two photosystems via state transitions (Tikkanen et al. 2006). The high rate of tandem duplication in the major *Lhc* gene families has been suggested to facilitate tuning of light harvesting to different light conditions (Cannon et al. 2004). The major Lhc proteins are only peripherally associated with the photosystem protein complexes. The less intimate association with the other subunits of the photosystems, compared to the minor Lhc proteins, might reduce dosage balance constraints (Veitia 2005), consistent with the higher rate of NP duplication in the major Lhc observed in several species.

In conclusion, we suggest that the photosystem protein complexes are dosage sensitive, which leads to retention of polyploid duplicates, and active elimination of non-polyploid duplicates, via purifying selection to maintain gene balance. Over time, balance of gene products is increasingly achieved via regulation of expression and becomes decoupled from gene dosage. This relaxes selection on gene copy number. Because the photosystems are highly functionally constrained, there are few opportunities for sub- or neofunctionalization, and most of the “extra” gene copies then begin to decay. We see remnants of this process in silenced *PsbO* and *PsbP* homoeologues in *Arabidopsis*. The CC, in contrast, is not dosage sensitive. Thus, redundant gene copies are neither actively eliminated nor maintained by selection, regardless of duplication mechanism (polyploidy or small-scale processes). These genes follow typical decay curves (Lynch and Conery 2000; Blanc and Wolfe 2004a; Schlueter et al. 2004; Maere et al. 2005), with most eventually being non-

functionalized. However, because CC genes potentially participate in multiple biochemical pathways, opportunities for functional differentiation (and long-term retention) are greater than for PSII or PSI. This, in turn, is manifested in older and larger gene families.

The fact that individual photosynthetic gene families do not exhibit consistent patterns of retention or loss across the three species examined here, or across nested polyploidy events within *Arabidopsis* (Figures 2.3 and 2.5), might seem to argue against the conclusion that higher-level functional groups are shaped by specific evolutionary forces. However, random mutational processes are likely to be driving both retention of dosage-insensitive CC gene duplicates (by facilitating functional divergence) and loss of dosage-sensitive PSII duplicates (by decoupling gene dosage from the amount of gene product). Thus, dosage sensitivity could produce a high overall rate of retention of polyploid duplicates in PSII, for example, despite individual gene families escaping this selective pressure by mutations that break the linkage between gene dosage and the abundance of gene product. Because these mutations are random, different gene families fractionate in different species, or following different polyploidy events in the same species.

The abundance of genomics studies observing trends in retention might give an exaggerated sense of consistency in terms of how particular genes respond to duplication. At least in the case of photosynthetic genes, such patterns dissolve when looking at the level of individual gene families, serving as a reminder that genome-level patterns are tendencies and not absolutes. Additionally, most studies looking at the behavior of broad functional classes of genes are restricted to a few species. Barker et al. (2008) found very different patterns of retention following polyploidy in the Compositae than have been observed in *Arabidopsis*, suggesting that duplicate gene evolution following polyploidy may follow family-specific trajectories. Additional

studies like the present one will help to reveal the extent to which “omics”-level patterns carry through to individual gene families, and to what extent patterns observed in one species can be extended to other species.

It should be noted that this study differs from previous genomics-level studies of polyploidy in that it investigates gene families in their specific physiological contexts. The reciprocal pattern of duplicate retention observed here, between the photosystems on one hand and the CC and LHC on the other, would not be detected when grouping genes by the functional categories used in these earlier studies, such as protein domains (Paterson et al. 2006), or gene ontologies (e.g., Blanc and Wolfe 2004b; Maere et al. 2005). The enzymes in a biochemical pathway (e.g., the CC), or subunits of a protein complex (e.g., PSII) are generally not characterized by common protein domains, and there are sufficient inconsistencies in GO annotations to preclude effective analysis of biochemical pathways and/or protein complexes via gene ontologies. For example, though there is a GO cellular component category for PSII (GO: 000953), this GO term has not been assigned to the *Arabidopsis* genes encoding three of the nine PSII subunits (PsbS, PsbW and PsbX). Similarly, there is a GO biological process term for “Carbon fixation” (GO:0015977), but gene families encoding four of 11 CC enzymes are not associated with this GO term. Thus, additional studies guided specifically by physiological or biochemical context will provide a valuable complement to existing studies using more generically assigned functional classifications.

## **METHODS**

*Tentative Consensus (TC)-based analyses:* Protein sequences were obtained from The Arabidopsis Information Resource (TAIR) website (<http://www.arabidopsis.org>) for all *Arabidopsis* genes involved directly in the Calvin

cycle (CC), photosystems I and II (PSI and PSII), and the light harvesting complexes (LHC). *Arabidopsis* protein sequences were used to query the *Glycine max* (release 12.0) and *Medicago truncatula* (release 8.0) gene indices maintained by the Dana Farber Cancer Institute (<http://compbio.dfci.harvard.edu/tgi/plant.html>) using TBLASTN. Sequences for all tentative consensus (TC) BLAST hits and corresponding *Arabidopsis* CDS sequences were translated to protein sequence and aligned using ClustalW, with default parameters, in BioEdit. Alignments were adjusted by eye as necessary. Singleton EST BLAST hits were excluded from analysis due to the frequency of errors in single EST sequences.

Gene phylogenies were constructed from the aligned sequences using maximum parsimony, as implemented in PAUP 4.0 (Swofford 2003). For alignments with fewer than 12 genes (including *Arabidopsis*, *Glycine* and *Medicago*), a full branch-and-bound search was performed. For alignments with 12 or more genes, a heuristic search was performed, with 1000 random addition sequence replicates, using Tree-Bisection-Reconnection branch swapping. To estimate divergences among *Glycine* and/or *Medicago* genes, the number of synonymous substitutions per synonymous site ( $K_s$ ) was calculated for each paralogous gene pair by the method of Yang and Nielsen, as implemented in PAML (Yang 1997) from the sequence alignments used to construct gene phylogenies.  $K_s$  values for pairs of *Arabidopsis* genes were taken from Blanc and Wolfe (2004a), or calculated as with *Glycine* and *Medicago*.  $K_s$  values were averaged for nodes joining more than two genes.

Duplications in *Glycine* and *Medicago* were categorized as resulting from polyploidy or non-polyploid (NP) duplication based on  $K_s$  values and gene tree topology as follows. The ca. 50 MY duplication event (hereafter referred to as “B”) occurred in the common ancestor of *Medicago* and *Glycine* (Pfeil et al. 2005; Figure 2.1). The median  $K_s$  value for this duplication event is 0.54 (0.40 to 0.72; +/- 1 SD)

(Schlueter et al. 2004). The median  $K_s$  value for the *Glycine*-specific, ca. 10MY duplication (hereafter referred to as “A”) is 0.12 (0.08 – 0.19) (A. Eagan pers. comm.; JAS unpublished data). A  $K_s$  value within either of these ranges for a pair of *Glycine* genes or within the older range for a pair of *Medicago* genes was taken as evidence of duplication by polyploidy (homoeology). Because the B duplication was shared by *Glycine* and *Medicago*, a *Medicago* sequence is expected to be sister to each *Glycine* lineage descended from this duplication. Because the A polyploidy was *Glycine*-specific, no *Medicago* sequence should nest within *Glycine* lineages resulting from this duplication. Gene phylogenies with this expected topology were considered further evidence for homoeology. Duplicate sequences were identified as homoeologues if they were supported by both  $K_s$  and gene tree topology. Due to the frequency of gene losses, duplicates in *Medicago* or *Glycine* were also considered homoeologues if supported by  $K_s$  even if gene losses in the other species had to be inferred. Duplicates were also considered homoeologues if  $K_s$  was outside of the range for that polyploidy event, but within 0.1 (B) or 0.02 (A) of the confidence interval for the polyploidy, and rejecting the event increased the number of losses inferred.

*Genomic synteny-based analyses:* *Arabidopsis* protein sequences for photosynthetic genes were used to query the *Glycine max* genome sequence (Glyma1 assembly; <http://www.phytozome.net/soybean.php>) and release 2.0 of the *Medicago truncatula* genome sequence (<http://www.medicago.org/genome/downloads/Mt2>) using TBLASTN. TC sequences identified through the TC-based analyses were also used to search the respective genome sequences using BLASTN. The CDS sequences of *Arabidopsis* and all *Glycine* loci showing significant BLAST scores ( $< 1e-5$ ) were aligned using CLUSTALW in BioEdit with default parameters.  $K_s$  and  $\omega$  ( $K_a/K_s$ ) were calculated by the method of Yang and Neilson, as implemented in PAML (Yang,

1997). Gene trees were constructed following the same methods used with the TC sequences.

In the absence of subsequent rearrangements, genes duplicated by polyploidy should reside in syntenic blocks (Zhang et al., 2002; Blanc et al., 2003; Bowers et al., 2003). Putative homoeologous blocks within the *Glycine max* genome (Glyma1 assembly; <http://www.phytozome.net/soybean.php>) were identified as follows. Gene families of sizes two to six were identified using Vmatch (<http://www.vmatch.de>). Pairwise matches were used as input for i-ADHoRE (Simillion et al. 2008). Corresponding soybean gene models (Glyma) were identified by BLASTN and homoeologous blocks pulled from the i-ADHoRE analysis.

We then determined whether each *Glycine* photosynthetic gene resides in or near (within 500 genes of) a synteny block. Pairs of gene family members residing within syntenic blocks were designated homoeologues.  $K_s$  estimates were used to determine from which polyploidy event (B or A) homoeologues were derived. For gene pairs close to, but not within, syntenic blocks, we manually searched for evidence of local synteny in a region of approximately 200 Kb centered on each gene using the Phytozome soybean genome browser (<http://www.phytozome.net/cgi-bin/gbrowse/soybean/>). Gene pairs within 500 genes of a synteny block that showed evidence for local synteny (at least three additional homologous gene pairs within 200 Kb) were also designated homoeologues. For genes not residing in or near syntenic blocks, we concluded that their homoeologues have been lost. Duplicate gene pairs that were not assigned to the B or A polyploidy events were assigned to one of three non-polyploid (NP) bins: pre-B, B-A, or A-present, based on  $K_s$  and gene tree topology.

For *Arabidopsis*, duplications resulting from the two most recent polyploidy events (designated “ $\beta$ ” and “ $\alpha$ ” by Bowers et al. 2003; Figure 2.1) were identified

previously by Blanc et al. (2003), Bowers et al. (2003) and Thomas et al. (2006;  $\alpha$  duplication only) using combinations of genomic synteny information, comparative phylogenetics and estimates of sequence divergence ( $K_s$ ). Lists of homoeologues are available at: <http://wolfe.gen.tcd.ie/blanc/supp/functional.html> (Blanc et al. 2003; Bowers et al. 2003) and <http://genome.cshlp.org/content/16/7/934/suppl/DC1> (Thomas et al. 2006). We searched these lists in order to identify homoeologues within the photosynthetic gene families investigated here. As with *Glycine*, gene pairs not identified as homoeologues were assigned to one of three NP duplication bins (pre- $\beta$ ,  $\beta$ - $\alpha$ , or  $\alpha$ -present) based on  $K_s$  and gene tree topology.

For all but three pairs of photosynthetic genes, the Blanc et al. (2003) and Bowers et al. (2003) datasets were consistent. The datasets differ for one pair each of *PGK*, *PsbTn* and *LhcB4*, and these discrepancies were resolved as described as follows.

For the *PGK* pair, Blanc et al. (2003) identified AT1G79550 and AT3G12780 as  $\beta$  homoeologues and Bowers et al. (2003) did not (Thomas et al. (2006) only identified  $\alpha$  homoeologues). We constructed a phylogeny from the Phytozome protein alignment for *PGK* (Viridiplantae gene cluster 9810655; <http://www.phytozome.net/>) using parsimony with default parameters in BioEdit. The protein phylogeny indicates that this duplication predates the monocot/dicot split (each *Arabidopsis* gene is sister to an *Oryza sativa* gene), so we concluded that this is a pre- $\beta$  gene duplication. These genes collapse into a gene cluster (N02922) generated by Tang et al. (2008) that represents an inferred ancestral gene duplicated by the ancient hexaploidy event. Thus, these genes are most likely the result of this hexaploidy event.

For the *PsbTn* gene pair, Bowers et al. (2003) and Thomas et al. (2006) identified AT3G21055 and AT1G51400 as  $\alpha$  homoeologues and Blanc et al. (2003) did not. This gene pair is contained within a large, well-conserved synteny block

identified by Blanc et al. (2003) (Block 0103319703610; <http://wolfe.gen.tcd.ie/athal/index.html>), suggesting that it was missed in their analysis, perhaps because the genes are short (103-106 amino acids), or because one of the two sequences overlaps with other gene models. Therefore, we concluded that these genes are indeed homoeologues from the  $\alpha$  polyploidy event.

Similarly, for the *LhcB4* gene pair, Bowers et al. (2003) and Thomas et al. (2006) identified AT3G08940 and AT5G01530 as  $\alpha$  homoeologues and Blanc et al. (2003) did not. As with *PsbTn*, these genes reside in a large, well-conserved synteny block (Block 0305069303260) in the Blanc et al. (2003) analysis, so we concluded that these are, indeed, homoeologues from the  $\alpha$  polyploidy.

For each photosynthetic gene family, we quantified the contributions of the various duplication events to each gene family using two parameters: percent retention and percent expansion (Figure 2.2). Percent retention for a given whole genome duplication was calculated by dividing the number of gene lineages duplicated by the WGD that are retained in duplicate today by the total number of gene lineages duplicated by the WGD. Percent expansion was calculated by dividing the number of gene lineages added by the given duplication event ( $\beta$ /B,  $\alpha$ /A, or NP) by the total number of gene lineages added since immediately before the  $\beta$ /B event. Mean percent retention and percent expansion for each function group (CC, PSII, PSI, or LHC) were calculated by averaging the values obtained for each gene family assigned to that functional group. This method weights each gene family equally, regardless of size.

Overall percent retention was also calculated for all photosynthetic genes combined, and for each functional group separately, in each of two ways. First, the number of lineages retaining duplicates from the specified duplication was divided by the total number of lineages present immediately prior to that duplication. Second, the number of genes present today that retain a homoeologue from the specified

duplication was divided by the total number of genes present today. Both methods differ from the mean percent retention method in that they effectively weight large gene families more heavily than small gene families. Of the two methods to calculate overall percent retention, the second method is comparable to the methods used by Schmutz et al. (2010), and was used for comparison to genome-wide retention estimates. This method yields higher estimates than the first because each retained homoeologue pair counts as two whereas singletons count only as one in both numerator and denominator.

*Tests of selection:* Global  $K_a/K_s$  values ( $\omega$ ) were calculated for all pairwise combinations of gene family members in *Arabidopsis* and *Glycine* using the method of Yang and Nielsen as implemented in PAML (Yang 1997). Sliding window  $K_a/K_s$  calculations were performed on homoeologue pairs from the recent polyploidy events in *Arabidopsis* and *Glycine* ( $\alpha$  and A, respectively) using the web tool, Sliding Window Analysis of  $K_a$  and  $K_s$  (SWAKK; Liang et al. 2006). We only analyzed duplicates from the  $\alpha$  and A duplications because these were the only duplication bins for which all functional groups (CC, PSII, PSI, LHC) have retained duplicates in both species. For three dimensional analyses, Protein Data Bank (PDB) files were obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) PDB website (<http://www.rcsb.org/pdb/>). We used the default window sizes of 30 amino acids (1D) and 10Å (3D).

*Expression analyses:* Correlation coefficients for photosynthetic genes in *Glycine* were calculated from 15 RNA-Seq experiments (cDNA libraries deep sequenced on the Illumina/Solexa platform) (Bolon et al. 2010, Libault et al. 2010). Tissue sources were as follows: four developmental stages of seed (25-50mg, 50-100mg, 100-200mg, and 200-300mg) from one low-protein near isogenic line (NIL) and one high-protein NIL (Bolon et al. 2010), root tips from 3-day old seedlings, roots

from 18-day old plants, root nodules collected 32 days after *B. japonicum* inoculation, leaves from 18-day old plants, apical meristems, open flowers, and 2-3 cm green seed pods (Libault et al. 2010).

For all photosynthetic genes in *Arabidopsis*, pairwise Pearson correlation coefficients ( $r$ ) were calculated from publically available microarray data using the web tool, CressExpress (<http://www.cressexpress.org/index.html>; Srinivasasainagendra et al. 2008), with default settings, and including all available tissue types and experiments.

## **ACKNOWLEDGEMENTS**

We thank Gary Stacey, Yung-Tsi Bolon, Bindu Joseph, Steven Cannon, and Michelle Graham for early access to their soybean RNA-seq datasets. We thank Tom Owens, Steven Cannon, and all members of the Doyle lab for helpful critiques of the manuscript. We acknowledge support from National Science Foundation grants IOS-0744306 and DEB-0709965.

## REFERENCES

- Adams KL, Wendel JF. 2005. Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids. *Genetics* 171: 2139-2142.
- Aury J, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444: 171-178.
- Baena-Gonzalez E, Aro E-M. 2002. Biogenesis, assembly and turnover of photosystem II units. *Philos T Roy Soc B*. 357: 1451-1460.
- Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore W, Knapp SJ, Rieseberg LH. 2008. Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years. *Mol Biol Evol* 25: 2445-2455.
- Birchler JA, Veitia RA. 2007. The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. *Plant Cell* 19: 395-402.
- Birchler JA, Yao H, Chudalayandi S. 2007. Biological consequences of dosage dependent gene regulatory systems. *Biochim Biophys Acta - Gene Structure and Expression* 1769: 422-428.
- Birchler JA, Veitia R. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol* 186: 54-62.
- Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* 13: 137-144.
- Blanc G, Wolfe KH. 2004a. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16: 1667-1678.

- Blanc G, Wolfe KH. 2004b. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16: 1679-1691.
- Bolon Y-T, Bindu J, Cannon S, Graham M, Diers B, Farmer A, May G, Muehlbauer G, Specht J, Tu Z et al. 2010. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biol* 10: 41.
- Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication event. *Nature* 422: 433-438.
- Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol* 4:10 doi:10.1186/1471-2229-4-10
- Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16: 738-749.
- De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20: 591-597.
- Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454: 765.
- Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. 2004. Evolution of the perennial soybean polyploidy complex (*Glycine* subgenus *Glycine*): a study of contrasts. *Biol. J. Linn. Soc.* 82:583-597.
- Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary Genetics of Genome Merger and Doubling in Plants. *Annu Rev Genet* 42: 443-461.
- Doyle JJ, Egan AN. 2010. Dating the origins of polyploidy events. *New Phytol* 186: 73-85.

- Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc Natl Acad Sci* 106: 5737-5742.
- Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* 16: 805-814.
- Ha M, Kim E, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci* 106: 2295-2300.
- Harrison EP, Lloyd JC, Raines CA. 1996. The effect of reduced SBPase levels on leaf carbon metabolism. *J Exp Bot* 47: 1306.
- Hwang HJ, Nagarajan A, McClain A, Burnap RL. 2008. Assembly and Disassembly of the Photosystem II Manganese Cluster Reversibly Alters the Coupling of the Reaction Center with the Light-Harvesting Phycobilisome. *Biochemistry* 47: 9747-9755.
- Klimmek F, Sjödin A, Noutsos C, Leister D, Jansson S. 2006. Abundantly and Rarely Expressed Lhc Protein Genes Exhibit Distinct Regulation Patterns in Plants. *Plant Physiol* 140: 793-804.
- Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* 20: 287-290.
- Lefebvre S, Lawson T, Zakhleniuk OV, Lloyd JC, Raines CA. 2005. Increased sedoheptulose-1,7-bisphosphatase activity in transgenic tobacco plants stimulates photosynthesis and growth from an early stage in development. *Plant Physiol* 138: 451-460.
- Liang H, Plazonic KR, Chen J, Li W, Fernández A. 2008. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet* 4: e11.
- Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G. 2010. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J* 9999: 9999

- Lundin B, Hansson M, Schoefs B, Vener AV, Spetea C. 2007. The Arabidopsis PsbO2 protein regulates dephosphorylation and turnover of the photosystem II reaction centre D1 protein. *Plant J* 49: 528-539.
- Lynch M, Conery JS. 2000. The Evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155.
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci* 102: 5454-5459.
- Minagawa J, Takahashi Y. 2004. Structure, function and assembly of Photosystem II and its light-harvesting proteins. *Photosynthesis Res* 82: 241-263.
- Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991.
- Miyagawa Y, Ichihara K, Tamoi M, Shigeoka S. 2001. Analysis of carbon metabolism in source and sink organs of transgenic tobacco plant having cyanobacterial FBP/SBPase in chloroplasts or cytosol. *Plant and Cell Physiol* 42: s172.
- Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424: 194-197.
- Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends Genet* 22: 597-602.
- Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ. 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. *Syst Biol* 54: 441-454.
- Sawchuk MG, Donner TJ, Head P, Scarpella E. 2008. Unique and overlapping expression patterns among members of photosynthesis-associated nuclear gene families in Arabidopsis. *Plant Physiol* 148: 1908-1924.

- Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868-876.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al. 2010. Genome sequence of the paleopolyploid soybean. *Nature* 463: 178-183.
- Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 18: 1152-1165.
- Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet* 20: 461-464.
- Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Curr Opin Plant Biol* 9: 104-109.
- Simillion C, Janssens K, Sterck L, Van de Peer Y. 2008. i-ADHoRe 2.0: an improved tool to detect degenerate genomic homology using genomic profiles. *Bioinformatics* 24: 127-128.
- Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot* 96: 336-348.
- Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE. 2008. CressExpress: a tool for large-scale mining of expression data from *Arabidopsis*. *Plant Physiol* 147: 1004-1016.
- Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van De Peer Y. 2005. EST data suggest that poplar is an ancient polyploid. *New Phytol* 167: 165-170.
- Sun N, Ma L, Pan D, Zhao H, Deng XW. 2003. Evaluation of light regulatory potential of Calvin cycle steps based on large-scale gene expression profiling data. *Plant Mol Biol* 53: 467-478.

- Swofford DL. 2003. PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4.
- Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18: 1944-1954.
- Thomas BC, Pedersen B, 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.
- Tikkanen M, Piipo M, Suorsa M, Sirpio S, Mulo P, Vainonen J, Vener AV, Allahverdiyeva Y, Aro EM. 2006. State transitions revisited – a buffering system for dynamic low light acclimation of *Arabidopsis*. *Plant Mol Biol* 62: 779-793.
- Tobin AK, Bowsher CG. 2005. Nitrogen and carbon metabolism in plastids: evolution, integration, and coordination with reactions in the cytosol. *Adv Bot Res* 42: 113-165.
- Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ et al. 2006. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* 18: 1348-1359.
- Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596-1604.
- Veitia RA. 2005. Gene dosage balance: deletions, duplications and dominance. *Trends Genet* 21: 33-35.
- Warner DA, Edwards GE. 1993. Effects of polyploidy on photosynthesis. *Photosynthesis Res* 35: 135-147.
- Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploidy speciation in plants. *Proc Natl Acad Sci* 106: 13875-13879.

- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555-556.
- Zhang Y, Xu GH, Guo XY, Fan LJ. 2005. Two ancient rounds of polyploidy in rice genome. *J Zhejiang Univ Sci B* 6: 87-90.
- Zhang L, Vision TJ, Gaut BS. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol Biol Evol* 19: 1464-1473.

## CHAPTER 3

# QUANTIFYING WHOLE TRANSCRIPTOME SIZE, A PREREQUISITE FOR UNDERSTANDING TRANSCRIPTOME EVOLUTION ACROSS SPECIES: AN EXAMPLE FROM A PLANT ALLOPOLYPLOID<sup>1</sup>

### ABSTRACT

Evolutionary biologists are increasingly comparing gene expression patterns across species. Due to the way in which expression assays are normalized, such studies provide no direct information about expression per gene copy (dosage responses) or per cell, and can give a misleading picture of genes that are differentially expressed. We describe an assay for estimating relative expression per cell. When used in conjunction with transcript profiling data, it is possible to compare the sizes of whole transcriptomes, which in turn makes it possible to compare expression per cell for each gene in the transcript profiling dataset. We applied this approach, using qRT-PCR and high throughput RNA sequencing (RNA-Seq), to a recently formed allopolyploid, and showed that its leaf transcriptome was approximately 1.4-fold larger than either progenitor transcriptome (70% of the sum of the progenitor transcriptomes). In contrast, the allopolyploid genome is 94.3% as large as the sum of its progenitor genomes, and retains  $\geq 93.5\%$  of the sum of its progenitor gene complements. Thus “transcriptome downsizing” is greater than genome downsizing. Using this transcriptome size estimate we inferred dosage responses for several

---

<sup>1</sup> This chapter has been published: Coate J.E. and Doyle J.J. 2010. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. *Genome Biology and Evolution* doi: 10.1093/gbe/evq038. Supplementary material cited herein is available at <http://gbe.oxfordjournals.org/>.

thousand genes and showed that the majority exhibit partial dosage compensation. Homoeologue silencing is non-randomly distributed across dosage responses, with genes showing extreme responses in either direction significantly more likely to have a silent homoeologue. This experimental approach will add value to transcript profiling experiments involving inter-species and inter-ploidy comparisons by converting expression per transcriptome to expression per genome, eliminating the need for assumptions about transcriptome size.

## **INTRODUCTION**

A growing number of transcript profiling studies, primarily using microarrays, have compared global expression patterns among closely related species, providing insights into a range of important evolutionary questions. Included among these are studies characterizing the selection pressures acting on gene expression in primates (Enard et al. 2002; Gilad et al. 2006), studies quantifying gene expression variation within and between populations or species of teleost fishes (Oleksiak et al. 2002), fruit flies (Rifkin et al. 2003), fungi (Andersen et al. 2008), and plants (Hammond et al. 2006), and several studies examining the effects of hybridization and genome doubling on gene expression in plants (Hegarty et al. 2005; Hegarty et al. 2006; Udall et al. 2006; Wang et al. 2006a; Flagel et al. 2008; Hegarty et al. 2008; Hovav et al. 2008a; Hovav et al. 2008b; Rapp et al. 2009). The advent of next generation sequencing technologies is likely to accelerate further the increase in such studies by removing many of the challenges associated with microarrays for inter-species comparisons (Gilad and Borevitz 2006; Blencowe et al. 2009; Gilad et al. 2009; Rokas and Abbot 2009).

Transcript profiling studies provide information about the relative abundances of transcripts. These and other expression assays such as RT-PCR and RNA blots

require normalization to correct for differences in amount of RNA template, as well as for other technical biases (Theillin et al. 1999; Quackenbush 2002), before comparisons can be made between samples. One or a few housekeeping genes are typically used as loading controls for RNA blots and RT-PCR assays, on the assumption that these genes are stably expressed across samples, thereby indicating the total amount of RNA used. With microarrays, raw data are generally normalized to total signal intensity (Quackenbush 2002) on the assumption that if the features on the array are a complete or unbiased sampling of the transcriptome, total signal intensity is a reasonable proxy for the whole transcriptome. For RNA-Seq data, read counts per gene are typically divided by gene length and total read count per sample (expressed as reads per kilobase per million; RPKM) to achieve comparable normalization (Marioni et al. 2008; Mortazavi et al. 2008). Consequently, for each of these assays, apparent differences in the expression of a gene between two samples are actually differences in expression per unit of RNA, or *per transcriptome* (Kanno et al. 2006).

Without information about the sizes of the two transcriptomes being compared, no inferences can be drawn from transcriptome-normalized expression about expression per gene copy, or expression per cell (Figure 3.1). Any difference in expression per cell between two samples that is proportional to the change in total transcriptome size will appear as equal expression per transcriptome. For example, in comparing a tetraploid with a diploid progenitor, genes showing equal expression per transcriptome (combining expression from the two homoeologous copies in the case of the tetraploid; Figure 3.1) could have equal numbers of transcripts per cell (if the transcriptomes are of equal size), or there could be twice as many transcripts per cell in the polyploid (if the polyploid transcriptome is doubled in size relative to the diploid; Figure 3.1). Conversely, genes exhibiting repression in the polyploid on a per

transcriptome basis could be expressed at an equal or even greater level per cell, again depending on the relative sizes of the two transcriptomes.

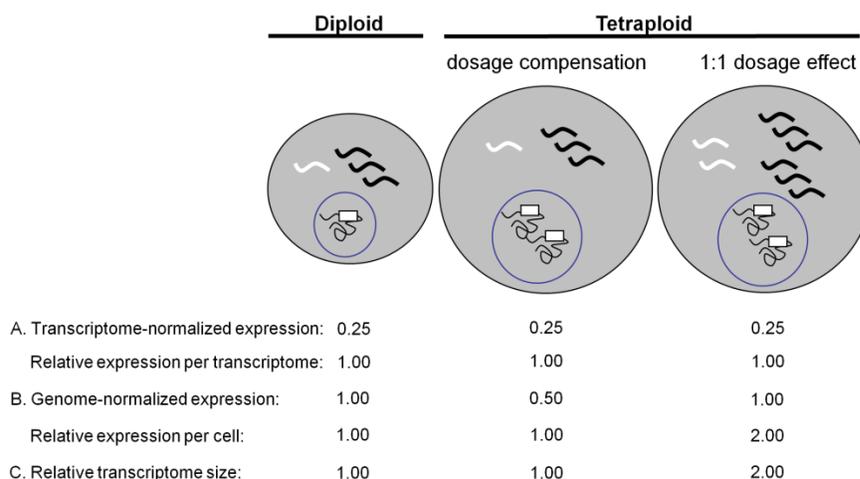


Figure 3.1. A comparison of transcriptome-normalized expression data vs. genome-normalized expression data. Grey circles represent cells, and wavy lines represent transcripts, with the diploid cell having a total of four transcripts in its transcriptome. Black circles represent nuclei, squiggly lines represent genomic DNA, and white boxes represent the genes encoding the white transcripts. A) *Transcriptome-normalized expression*. Expression of the white transcript, measured on a per transcriptome basis, is 0.25 (1 transcript out of a total of 4 transcripts) in the diploid. The same transcriptome-normalized expression values are obtained in two tetraploids showing different expression levels per cell, illustrating that transcriptome-normalized measurements do not provide information on transcript abundance per genome (dosage response), or per cell. B) *Genome-normalized expression*. If the expression of the white transcript is instead normalized to genome copy number (1 for the diploid, 2 for the tetraploid), differences in transcript abundance per cell become apparent, and dosage responses can be determined. Relative expression per cell in the tetraploid is simply two times the genome-normalized expression. C) *Relative transcriptome size*. Tetraploid transcriptome size (relative to the diploid transcriptome) can then be estimated by dividing relative expression per cell by relative expression per transcriptome.

The unstated assumption of expression studies is that the transcriptomes being compared are of equal size. This seems an unwarranted assumption, particularly when

comparing polyploids and diploids, because transcriptome sizes are likely to differ due to genome-wide differences in gene dosage. But even for comparisons not involving ploidy differences, the potential exists for transcriptome sizes to differ, especially when comparisons are made across tissue types, developmental stages or species, for which microarray experiments frequently observe dramatic differences in transcriptome-normalized expression profiles (Hammond et al. 2006; Andersen et al. 2008). Given numerous differences in transcriptome-normalized expression, what is the net effect on transcriptome size? In the absence of a method to quantify this effect, it is not possible to determine what such differences, at the level of individual genes, mean in terms of transcript abundance per cell. Thus, a method to estimate relative transcriptome sizes is needed for determining expression differences per cell based on transcriptome-normalized expression profiling data. This is particularly true for expression studies of polyploids.

Most, if not all, flowering plants have experienced one or more whole genome duplications (polyploidy events) during their evolutionary histories (Cui et al. 2006; Tang et al. 2008), and an estimated 15% of angiosperm speciation events are associated with increases in ploidy (Wood et al. 2009). Polyploids often appear to be more successful than their diploid progenitors, as measured by broader geographical ranges (Ehrendorfer 1980; Otto and Whitton 2000), and greater capacity to tolerate stressful environments (Stebbins 1971; Lewis 1980; Grant 1981; Otto and Whitton 2000; Hegarty and Hiscock 2008), and it has been proposed that polyploidy contributed to the survival of several plant lineages through the Cretaceous-Tertiary mass extinction (Fawcett et al. 2009).

Changes in gene expression, due to epigenetic mechanisms, transposon activation, sequence changes, novel combinations of regulatory factors and/or increased gene dosage, are thought to underlie this apparent success (Chen 2007).

Consequently, a central focus of polyploidy research is in understanding transcriptional responses to genome duplication.

For every gene duplicated by polyploidy, a range of dosage responses (changes in expression associated with changes in gene dosage) is possible. The two most obvious are dosage compensation, in which expression is modulated to 1.0x diploid levels per cell, or 0.5x per genome; and 1:1 dosage effects, resulting in 2.0x diploid expression per cell, or 1.0x per genome. Other responses are also possible, including partial dosage compensation (expression between 1.0-2.0x diploid level per cell, or 0.5-1.0x per genome), negative dosage effects (expression <1.0x diploid level per cell, or <0.5x per genome), and >1:1 dosage effects (expression >2.0x diploid level per cell, or >1.0x per genome). “Dosage effect” and “dosage compensation” refer most clearly to comparisons of an artificial autopolyploid with the diploid genotype from which it was synthesized. In an allopolyploid that combines two differentiated diploid genomes, the situation is more complex. Additivity of the two parental expression levels for a given gene would be the equivalent of a 1:1 dosage effect, with midparent expression levels being analogous to dosage compensation. Regardless of the type of polyploidy involved, the cumulative effect of these dosage responses will dictate to what extent the polyploid transcriptome differs in size from its diploid progenitor transcriptome(s).

There is little information available about gene dosage responses following polyploid duplication. In a seminal investigation of a synthetic maize (*Zea mays*) autopolyploid series, Guo et al. (1996) established that rRNA exhibits a 1:1 dosage effect in response to changes in ploidy, then used rRNA as a loading control for Northern blots in order to determine dosage responses for 18 genes. Most of the 18 genes investigated exhibited a 1:1 dosage effect. There were, however, several exceptions, with some genes showing negative dosage effects, others showing >1:1

dosage effects, and others showing variable responses depending on the specific ploidy level (“odd/even effects”). Beyond this study, the literature is largely silent, with no equivalent data available for natural autopolyploids, or for natural or synthetic allopolyploids. Thus, it remains an open question how the responses observed by Guo et al. (1996) extend to other genes, other tissues and other species, and how the responses of individual genes sum over the transcriptome as a whole. There exists no literature on overall transcriptome size in polyploids relative to their diploid progenitor(s).

Here we have calculated the relative size of an allopolyploid leaf transcriptome by combining genome-normalized expression estimates from a novel qRT-PCR assay with transcriptome-normalized expression estimates from RNA-seq. By this approach we made seven independent measurements of relative transcriptome size, which we then used to test two hypotheses: 1) The allopolyploid transcriptome is equal in size to the midparent transcriptome (genome-wide dosage compensation); and 2) The tetraploid transcriptome is equal to the sum of its progenitor transcriptomes (a genome-wide dosage effect). We then used our estimate of transcriptome size to estimate expression per genome and per cell in the allopolyploid relative to its diploid progenitors, for ca. 15,000 genes in the RNA-Seq dataset. This made it possible to quantify the frequency distributions, as well as patterns of homoeologue deployment, for each kind of dosage response.

## **MATERIALS AND METHODS**

*Plant material:* The study group consisted of the natural allopolyploid, *Glycine dolichocarpa* ( $2n = 80$ ; designated “T2”) and its diploid progenitors, *G. tomentella* ( $2n = 40$ ; “D3”) and *G. syndetika* ( $2n = 40$ ; “D4”). (Doyle et al. 2004; Pfeil et al. 2006). The two diploid species, D3 and D4, diverged approximately 2.5 MYA, and

hybridized to give rise to T2 within the last 100,000 years (Doyle et al. 2004). T2 is therefore a fixed hybrid, whose genome comprises two homoeologous subgenomes, one contributed by D3 and the other by D4. Therefore at each locus in T2 there is a D3 and a D4 allele, except in cases where the D3 or D4 homoeologue has been lost during the relatively short time since the formation of T2.

Plants were grown in a common growth chamber with a 12hr/12hr light/dark cycle and  $125 \mu\text{mol m}^{-2} \text{s}^{-1}$  light intensity. Young, fully expanded leaflets were collected 1.5 – 2.0 hours into the light period and frozen in liquid nitrogen.

*Genome-Normalized Expression Assay:* In order to estimate relative expression level per genome, we devised a qRT-PCR assay that normalizes cDNA amplification to genomic DNA amplification. The key to this assay is simultaneously extracting both RNA and genomic DNA (gDNA) from the same tissue so that *in vivo* RNA/gDNA ratios are preserved. Primers that specifically amplify either cDNA or gDNA were then used for qRT-PCR, allowing for normalization of gene expression (cDNA amplification) to genome copy number (gDNA amplification). This contrasts with typical qRT-PCR assays, in which target cDNA amplification of a target gene is normalized to cDNA amplification of a reference gene.

Leaflets were pooled from six individuals for each biological replicate. Three biological replicates were analyzed per species. RNA and gDNA (total nucleic acid; TNA) were co-extracted from each biological replicate using the BioChain<sup>®</sup> Dr. P Isolation Kit, with the following modifications: 1) centrifugation steps were performed at room temperature. 2) The DNA/RNA pellet obtained from the isopropanol precipitation was washed 3x with 70% EtOH, then resuspended in DEPC H<sub>2</sub>O/0.1% EDTA. This TNA suspension was then used as the template for reverse transcription. RNA, in a mixture with gDNA (approx. 1 $\mu\text{g}$  TNA), was reverse transcribed with random decamers using the Ambion Retroscript kit.

Primers were designed to be specific to either cDNA or gDNA as follows. For cDNA-specific primers, one or both primers in a pair were designed to span exon-exon splice junctions so that they would not anneal to unspliced gDNA. For gDNA-specific primers, one or both primers were designed to prime at least partially within an intron so that they would not anneal to spliced cDNA. Template specificity was confirmed for all primer pairs by semi-quantitative PCR with cDNA and gDNA templates. Primer target sequences were confirmed for each gene in all three species by Sanger sequencing. Primers specific to cDNA were designed for seven genes or gene families (Table 1 and Supplementary Table 1). Primers specific to gDNA were designed to three genes or gene families (Supplementary Table 1).

Table 3.1. Genes and gene families for which expression was analyzed by genome-normalized qRT-PCR.

<i>G. max</i> Gene ID(s) <sup>1</sup>	Annotation <sup>2</sup>	Unique RPM <sup>3</sup> - T2	Transcriptome-normalized expression ratio	
			T2/D3	T2/D4
Glyma13g23150, Glyma17g11720	<i>MGD</i>	30.6	2.5	1.4
Glyma15g32540	<i>EMB1473</i>	271.6	1.1	0.7
Glyma04g39380, Glyma06g15520	<i>Actin</i>	425.3	1.1	1.4
Glyma18g03440, Glyma11g34900	<i>SBPase</i>	1260.4	1.0	1.0
Glyma13g32920	<i>Defense-related</i>	1315.2	7.4	3.3
Glyma04g42870, Glyma06g11890	<i>PsbS</i>	1903.2	0.8	1.2
Glyma05g00620	<i>PsaF</i>	8936.1	0.9	1.7

<sup>1</sup>*Glycine max* locus identifiers - <http://www.phytozome.net/cgi-bin/gbrowse/soybean/>  
Where two gene IDs are listed, cDNA-specific primers amplify both.

<sup>2</sup>MGD - Monogalactosyldiacylglycerol synthase; EMB1473 – Embryo defective 1473; SBPase – Sedoheptulose-1,7-bisphosphatase; PsbS – subunit S of photosystem II; PsaF – subunit F of photosystem I

<sup>3</sup>RPM = reads per million

The cDNA/gDNA mixture was diluted five-fold and used as template for qRT-PCR with the following components: 5.75ul H<sub>2</sub>O, 7.5ul Power SYBR Green master mix (Applied Biosystems), 0.375ul forward primer, 0.375ul reverse primer, and 1ul

template. Assays were performed on an Applied Biosystems 7900 HT instrument, with 40 PCR cycles. Dissociation curves were generated at the end of the PCR to confirm specificity of amplification. For each primer pair and species, we amplified three technical replicates from each of three biological replicates.

Amplification efficiencies were estimated using LinRegPCR (Ramakers et al. 2003) for each individual reaction. Mean efficiency per amplicon was used for relative expression estimates. Expression of each target gene (cDNA-specific amplification) was normalized to genome copy number, as estimated by the geometric mean of amplification from the three gDNA-specific targets. Relative genome-normalized expression values (T2/D3, T2/D4, T2/midparent, and D4/D3) were estimated using the Relative Expression Software Tool (REST) (Pfaffl et al. 2002).

We confirmed that T2 retains both D3 and D4 homoeologues for all gene targets (both cDNA and gDNA-specific) by the presence of both D3- and D4-specific SNPs, as revealed by sequence data from the transcript profiling experiment and/or from Sanger sequencing of cDNA and/or gDNA (Supplementary Figure S1 and Supplementary Table 2). Because T2 has twice as many copies of each target gene as the diploids, relative expression per cell in T2 (T2/D3, T2/D4 and T2/midparent) was obtained by multiplying relative expression per genome by two. For comparisons of D3 and D4, expression per genome is equivalent to expression per cell.

*Transcriptome-Normalized Expression Assay:* Relative expression per transcriptome was measured by RNA-Seq. Leaflets were pooled from six individuals per species, and RNA was isolated using the Qiagen Plant RNeasy kit with on-column DNase treatment. Sequencing was performed using Solexa/Illumina “Sequencing by Synthesis” with the following modifications. Poly A+ RNA was annealed to high concentrations of random hexamers, reverse transcribed, and ligated to adapters complementary to sequencing primers. The cDNA was then amplified by 20 cycles of

polymerase chain reaction and size fractionated on agarose gels. 200 bp amplicons were excised and sequenced by synthesis with reversible terminator nucleotides with cleavable fluorescence.

To process the data for analysis, files were mirrored to an off-instrument computer using the Illumina® platform to perform image analysis, base-calling, quality filtering, and per base confidence scores. Sequences were then aligned using GSNAP (Wu and Nacu 2010) against the 8X genome sequence of soybean (*Glycine max*; version Glyma1, Soybean Genome Project, DoE Joint Genome Institute), which diverged from the common ancestor of D3, D4 and T2 approximately 5 MYA (Innes et al. 2008). Note that soybean, D3, and D4, all of which are  $2n = 40$ , are fully diploidized descendants of an ancestor that underwent a whole genome duplication approximately 10MYA (Shoemaker et al. 2006). Roughly half of the genes duplicated by this event are retained in duplicate in the soybean genome (Schmutz et al. 2010). Only reads mapping unambiguously to a single copy in the soybean genome were used in this study.

GSNAP was parameterized to allow spliced alignments of the transcript reads to the genomic reference sequences requiring canonical splice sites and allowing introns of up to 10Kbp; alignments were also allowed to include small indels and mismatches, but required that at least 30 out of the 36 base pairs in a read were matched. Alignments above this threshold with the highest number of identities were divided into three classes: uniquely aligned reads, low-copy repetitive alignments matching no more than 5 locations in the reference and highly repetitive reads matching >5 locations in the reference. The alignments in the first two classes were further processed using the Alpheus pipeline (Miller et al. 2008) for deriving per-gene read counts and sequence polymorphism calls. The boundaries of each gene were taken as the maximal starting and ending positions from any of the transcripts

associated with the gene, and any read alignment partially contained within this span was counted toward the expression of that gene in the given sample. Reads from uniquely aligned sequences were used to estimate expression levels after normalizing read counts to account for overall sampling sizes. Transcript abundance per transcriptome for a given gene was estimated as the number of reads unambiguously mapped to that gene per million unambiguously mapped reads generated by that library (reads per million; RPM). Because all comparisons involved the relative expression of individual genes across species (as opposed to multiple genes within a species), no adjustment for gene length (e.g., RPKM) (Mortazavi et al. 2008) was necessary.

*Calculation of Relative Transcriptome Size:* We obtained independent estimates of relative transcriptome size (T2/D3, T2/D4, T2/midparent, and D4/D3) from each of the seven genes assayed by qRT-PCR. The expression per cell (qRT-PCR) estimate obtained for each gene was divided by expression per transcriptome (RNA-Seq) for that gene. The mean of these seven independent estimates (and associated standard error) was taken as the best overall estimate of relative transcriptome size.

*Comparison of cDNA Pools from Genome-Normalized and Transcriptome-Normalized Expression Assays:* One of the cDNA-specific primer pairs employed in the qRT-PCR assay amplifies two actin loci (Supplementary Table 1). In order to confirm that the RNA extracted with the Dr. P kit (used for the qRT-PCR assay) was comparable to the RNA extracted with the Qiagen RNeasy kit (used for RNA-Seq), and quantitatively representative of its corresponding transcriptome, expression of the other six genes assayed by qRT-PCR was also normalized to the combined expression of the actin genes, and relative expression ratios for T2 vs. each diploid estimated using REST, as above. RNA-Seq unique RPMs for each of the same six genes were

then normalized to the same two actin genes ( $\text{RPM}_{\text{target gene}} / \text{RPM}_{\text{actin}}$ ). The actin-normalized expression ratios from qRT-PCR were then compared to the actin-normalized ratios from RNA-Seq to determine the correlation of actin-normalized expression estimates between the two platforms (Supplementary Figure S2).

*Estimation of relative homoeologue expression levels in T2:* We checked each nucleotide position within exons for substitutional differences distinguishing D3 from D4 using consensus sequences from the Illumina reads. Only sites covered by at least two reads in both D3 and D4 were used. For each site that differed between D3 and D4, and to which we had aligned at least 5 reads from the T2 sample, we determined the proportion of D3-type versus D4-type nucleotides sampled. The homoeologue expression ratio for a gene was calculated by averaging the ratios at each diagnostic site weighted by the number of T2 reads aligned across that site.

*Estimation of Genome Sizes and Extent of Endopolyploidy:* Young, fully expanded leaflets were collected and stored overnight in the dark on wet paper towels. Leaves were finely chopped in an  $\text{MgSO}_4$  buffer (Arumuganathan and Earle 1991) and passed through a 30 micron mesh filter (Partec CellTrics) to remove large debris. Propidium iodide (15  $\mu\text{l}$  of a 5  $\mu\text{g}/\text{ul}$  solution) and RNase (5  $\mu\text{l}$  of a 5  $\text{mg}/\text{ml}$  solution) were then added to the filtrate. Samples were run on a Coulter Epics XL-MCL flow cytometer. Measurements of fluorescence intensity were made on 3-4 individuals per species. Data were analyzed using WinMDI.

Absolute genome sizes were estimated by co-chopping 12.5 mg of leaf tissue with 12.5 mg of leaf tissue from a plant standard of known genome size. *Glycine max* (2.5  $\text{pg}/2\text{C}$ ) and *Zea mays* (5.4  $\text{pg}/2\text{C}$ ) were used as standards for the tetraploid and diploids, respectively (Dolezel et al. 2007).

The extent of endoreduplication was estimated by analyzing 25mg of leaf tissue without an internal standard. Endoreduplication produces peaks in the

fluorescence histogram in multiples of the main (2C) peak. The ratio of endoreduplicated nuclei to total nuclei was quantified by dividing the number of nuclei in the endopolyploid peaks by the combined number of nuclei in the primary and endopolyploid peaks.

## RESULTS

*Expression per Genome:* We devised a novel qRT-PCR assay that utilizes genomic DNA (gDNA) and RNA co-extracted from the same tissue to normalize transcript abundance to gDNA abundance. Because RNA and gDNA were extracted from the same cells, *in vivo* RNA/gDNA ratios were preserved. In addition, we confirmed that amplification efficiencies were comparable ( $\geq 1.90$ ) in all three species for each primer pair used in the qRT-PCR assay (data not shown). Consequently, normalizing cDNA amplification by gDNA amplification in qRT-PCR gives a direct readout of transcript abundance per genome. Using this method, we quantified expression per genome in the allotetraploid (T2) and its diploid progenitors (D3 and D4) for seven different genes or gene families (Table 1). Across the seven genes/gene families, expression per genome in T2 relative to the midparent value ranged from 0.6x to 3.7x (Figure 3.2 and Supplementary Table 3).

Based on RNA-Seq (see below) and/or Sanger sequencing, we confirmed that T2 retains both D3 and D4 homoeologues for each target gene used in the qRT-PCR assay (Supplementary Figure S1 and Supplementary Table 2). Because T2 has two copies of each gene used for genomic normalization for every one copy in the diploids (two homoeologues per diploid gene), we calculated expression per *cell* in T2 relative to its diploid progenitors as two times the relative expression per genome (Supplementary Table 3).

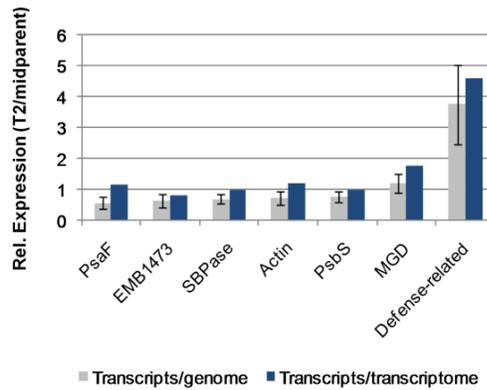


Figure 3.2. qRT-PCR based estimates of transcripts per genome (gray;  $\pm$  SE; N = 3) and RNA-Seq based estimates of transcripts per transcriptome (blue; N = 1) in T2 relative to the midparent values for seven genes or gene families. Values are ordered by relative expression per genome.

*Expression per Transcriptome:* We also profiled the leaf transcriptomes of the allotetraploid (T2) and its diploid progenitor species (D3 and D4) by RNA-Seq. High throughput sequencing using Solexa/Illumina technology generated > 5 million 36bp reads for each species. Reads were uniquely mapped to > 35,000 genes in each species, with unique read counts per gene ranging from 1 to > 98,000, reflecting the relative abundance of that transcript in the transcriptome (Marioni et al. 2008). The expression level per transcriptome for a given gene was estimated as the number of sequencing reads derived from that gene divided by the total number of reads derived from that sample, reported as reads per million (RPM). Because we compared the relative expression of individual genes across species (as opposed to multiple genes within a species), relative expression estimates were not affected by variation in gene length, making length adjustments (e.g., RPKM) (Mortazavi et al. 2008) unnecessary. Across the seven genes/gene families for which relative expression per *genome* was determined by qRT-PCR, expression per transcriptome in T2 relative to the midparent value ranged from 0.8x to 4.6x (Figure 3.2 and Supplementary Table 3).

*Comparison of cDNA Pools from Genome-Normalized and Transcriptome-Normalized Expression Assays:* Because the cDNA template used in the qRT-PCR assay was generated in a non-standard way (reverse transcription was performed on RNA in a native mixture with gDNA), we verified that these cDNA pools were quantitatively equivalent to the cDNA pools used for RNA-Seq. Following standard qRT-PCR methodology, expression estimates obtained using the TNA-derived cDNA for six of the seven genes examined were normalized to the expression of actin (the seventh gene family), and relative expression ratios for T2 vs. the diploid midparent value were estimated. RNA-Seq RPMs for each of the same six genes were then normalized to the same actin genes ( $\text{RPM}_{\text{target gene}} / \text{RPM}_{\text{actin}}$ ). The actin-normalized expression ratios from qRT-PCR were then compared to the actin-normalized expression ratios from RNA-Seq. Across the six genes, a strong correlation was observed between the two estimates (Pearson correlation coefficient = 0.99; Figure S2), indicating that the RNA-Seq and qRT-PCR cDNA preps were equivalently representative of the transcriptomes from which they were derived.

*Relative Transcriptome Size:* To estimate the size of the tetraploid transcriptome relative to each diploid transcriptome, we then divided the per cell expression ratios from the qPCR assay by the per transcriptome expression ratios from the RNA-Seq dataset (Figure 3.1). The logic of this calculation can be seen algebraically. The qPCR result gives the expression of a gene in the tetraploid relative to the expression in the diploid on a per cell basis (Figure 3.1):

$$\text{Ratio 1: } \frac{\frac{\text{Target gene transcripts}}{\text{cell}} (\text{tetraploid})}{\frac{\text{Target gene transcripts}}{\text{cell}} (\text{diploid})}$$

The RNA-Seq result gives the expression in the tetraploid relative to the expression in the diploid on a per transcriptome basis (Figure 3.1):

$$\text{Ratio 2: } \frac{\frac{\text{Target gene transcripts}}{\text{total transcripts}} (\text{tetraploid})}{\frac{\text{Target gene transcripts}}{\text{total transcripts}} (\text{diploid})}$$

Dividing ratio 1 by ratio 2 yields the following:

$$\text{Ratio 3: } \frac{\frac{\text{Total transcripts}}{\text{cell}} (\text{tetraploid})}{\frac{\text{Total transcripts}}{\text{cell}} (\text{diploid})}$$

This is the size of the tetraploid transcriptome relative to the size of the diploid transcriptome.

With this approach we obtained seven independent estimates of the size of the tetraploid transcriptome relative to each diploid progenitor transcriptome, and to the diploid midparent transcriptome (Figure 3.3A; Supplementary Table 3). There was variation among individual gene estimates, as might be expected given that there is error associated with both RNA-Seq and qPCR data, but all estimates for T2/midparent fell between 1- and 2-fold (the expected values if the T2 transcriptome overall was dosage compensated, or exhibited 1:1 dosage effects, respectively).

With these data, we rejected the null hypothesis that the T2 transcriptome was doubled (a genome-wide dosage effect) relative to the midparent transcriptome ( $p = 0.0002$ ; One-sample t-test), as well as the null hypothesis that the T2 transcriptome was equal in size (genome-wide dosage compensation) to the midparent transcriptome ( $p = 0.0031$ ; One-sample t-test). On a global scale, therefore, the T2 leaf transcriptome has been partially dosage compensated. Our data indicated that the leaf transcriptome

of the tetraploid under these conditions was 1.4-fold ( $\pm 0.1$  SE) larger than the midparent transcriptome (Figure 3.3B), and 1.3- to 1.4-fold ( $\pm 0.2$  SE) larger than the transcriptomes of either individual diploid progenitor. The diploid transcriptomes did not differ significantly in size ( $p = 0.7561$ ; one-sample t-test). We estimated that the D4 leaf transcriptome was 1.1-fold ( $\pm 0.2$  SE) larger than the D3 leaf transcriptome (Figure 3.3B).

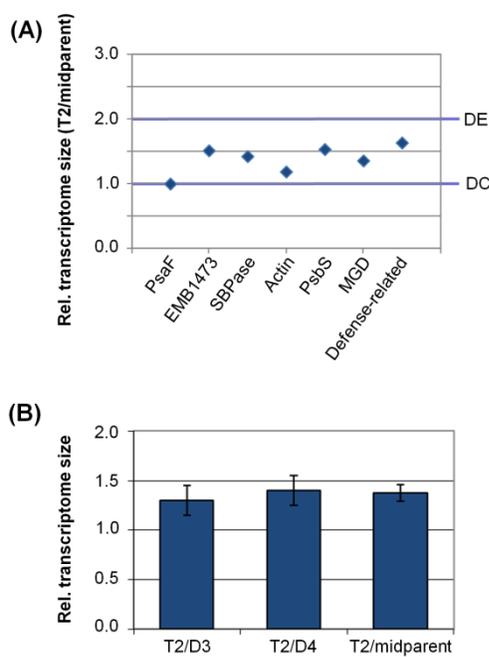


Figure 3.3. T2 transcriptome size relative to the transcriptomes of its diploid progenitors. A) Seven individual gene-based estimates of relative transcriptome size (T2 vs. the diploid midparent transcriptome). “DE” designates the expected value if the T2 transcriptome experienced genome-wide 1:1 dosage effects. “DC” designates the expected value if the T2 transcriptome experienced genome-wide dosage compensation. B) Average estimate of tetraploid transcriptome size relative to the transcriptomes of each diploid progenitor, and to the midparent diploid transcriptome ( $\pm$  SE;  $N = 7$ ).

Endopolyploidy (the occurrence of different ploidy levels within different cells of an organism) is common in seed plants (Barow 2006). Because our transcriptome

size estimates were obtained by normalizing gene expression to ploidy level (genome copy number), differences in the extent of endopolyploidy between T2 and D3 or D4 would affect our estimates of transcriptome size. In order to quantify the extent of endopolyploidy in D3, D4 and T2, we performed flow cytometry on leaf tissue of a comparable developmental stage (young, fully expanded) as was used for RNA-Seq and qRT-PCR. We observed minimal levels of endopolyploidy in all three species, with comparable fractions of endopolyploid nuclei in each (4-7% of nuclei; Supplementary Table 4 and Supplementary Figure S3). Our estimates of transcriptome size are not, therefore, skewed by differences in endopolyploidy.

*Dosage Responses Across the Tetraploid Transcriptome:* Once an estimate of transcriptome size was obtained, estimates of dosage response could then be made for each gene in the transcriptome profiling dataset. Because the T2 transcriptome was estimated to be 1.4-fold ( $\pm 0.1$  SE) larger than the midparent diploid transcriptome, a gene that has undergone complete dosage compensation in T2 would exhibit a transcriptome-normalized expression level of 0.7 times the midparent diploid level (0.7x diploid copies per transcriptome  $\times$  1.4 diploid transcriptome equivalents per cell  $\approx$  1.0x diploid copies per cell, or 0.5x copies per genome). Likewise, a gene whose expression has experienced a 1:1 dosage effect would exhibit a transcriptome-normalized expression level of 1.4x the midparent level (1.4  $\times$  1.4  $\approx$  2.0x copies per cell, or 1.0x copies per genome). Based on the SE associated with our estimate of transcriptome size ( $\pm 0.1$ ), a 95% confidence interval for the size of the T2 transcriptome relative to the midparent value is approximately 1.2 to 1.6-fold (1.4  $\pm$  1.96  $\times$  SE). From this, we approximated confidence intervals for each response: genes exhibiting transcriptome-normalized expression in T2 between 0.6 – 0.8x the midparent level were most likely dosage compensated (0.6  $\times$  1.6  $\approx$  1.0; 0.8  $\times$  1.2  $\approx$  1.0), and genes exhibiting transcriptome-normalized expression between 1.3-1.7x the

midparent level most likely exhibited a 1:1 dosage effect ( $1.3 \times 1.6 \approx 2.0$ ;  $1.7 \times 1.2 \approx 2.0$ ).

Figure 3.4A shows the distribution of dosage responses in T2. Of 15,761 genes in our RNA-Seq dataset with at least 10 uniquely mapped RPM in one of the three species, 2,319 (14.7%) exhibited transcriptome-normalized expression in T2 consistent with dosage compensation, and 2,724 (17.3%) exhibited expression levels consistent with a 1:1 dosage effect. The majority of genes in T2 (8,115; 51.5%) displayed an intermediate dosage response (0.8-1.3x midparent). Of the remaining genes, 1,583 genes (10.0%) exhibited a negative dosage effect ( $<0.5x$  diploid expression per genome), and 1,020 genes (6.5%) exhibited a greater than 1:1 dosage effect ( $>1x$  diploid expression per genome).

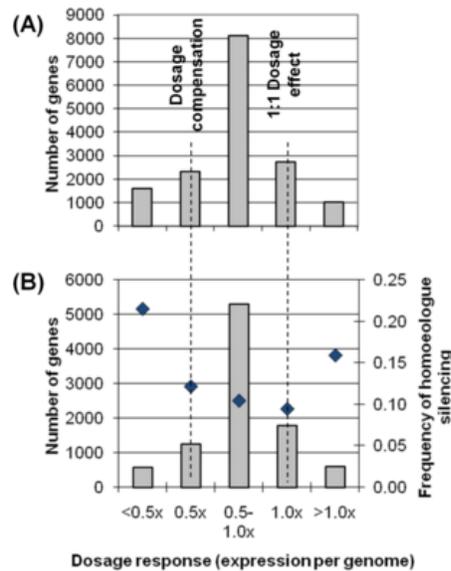


Figure 3.4. Genome-wide distribution of gene dosage responses and homoeologue silencing in the T2 allotetraploid. A) Number of genes from the RNA-Seq dataset with  $\geq 10$  unique RPM in at least one of the three species showing specified dosage responses in T2. B) Of the genes from panel A for which homoeologue expression could be estimated, number of genes showing specified dosage responses (grey bars), and the fractions of each for which one homoeologue is silent ( $\blacklozenge$ ). Dosage responses are expressed as relative expression per genome (T2/midparent).

*Homoeologue Modulation:* RNA-Seq enables estimates of the contributions of each homoeologue to total expression in the T2 tetraploid (see Figure S1). Combining this information with our estimates of dosage response, we could identify patterns of homoeologue deployment associated with each dosage response category. For example, dosage compensation could be achieved by silencing one of two homoeologues while maintaining the other at its diploid expression level, or by down regulating both copies. Of the 2,319 genes that were considered to be dosage compensated, homoeologue contributions could be determined for 1,240. Of these 1,240 genes, 151 (12.2%) expressed only one homoeologue (Figure 3.4B). By comparison, 168/1,772 (9.5%) genes that exhibited a 1:1 dosage effect expressed only one of two homoeologues (Figure 3.4B). Thus, there was a slight but significant increase in the frequency of homoeologue silencing amongst dosage-compensated genes vs. genes that showed a 1:1 dosage effect ( $\chi_1^2 = 5.60$ ,  $p = 0.02$ ). Nonetheless, even amongst dosage-compensated genes, the vast majority expressed both homoeologues, indicating that in most cases dosage compensation was achieved by more subtle modulations of homoeologue expression.

As might be expected, genes that exhibited negative dosage effects ( $<0.5x$  diploid expression per genome) silenced homoeologues at the highest frequency (21.5% of 572 genes; Figure 3.4B), which was significantly higher than for genes that were dosage compensated ( $\chi_1^2 = 26.53$ ,  $p < 0.0001$ ). Surprisingly, the next highest category of genes with one silenced homoeologue was the group of genes showing  $>1:1$  dosage effects (15.9% of 603 genes), which was also significantly higher than the group of genes that were dosage compensated ( $\chi_1^2 = 4.90$ ,  $p = 0.03$ ). Thus, many loci showing strongly up-regulated expression in T2 vs. its diploid progenitors did so using only one of two homoeologues. Overall, a pattern emerged in which genes showing the most extreme dosage responses in either direction ( $<0.5x$  or  $>1.0x$  diploid

expression per genome) were more likely to exhibit homoeologue silencing than genes showing intermediate responses (0.5 – 1.0x diploid expression per genome;  $\chi_1^2 = 66.73$ ,  $p < 0.0001$ ; Figure 3.4B).

*Transcriptome vs. Genome Size:* Using flow cytometry, we estimated the T2 genome to be 1.89-fold larger than the midparent genome (1.84-fold larger than D3 and 1.93-fold larger than D4), or 94.5% of the sum of the two progenitor genomes (Supplementary Table 4). Of 10,311 genes with sufficient depth of sequence coverage in the RNA-Seq dataset, and diagnostic SNPs distinguishing D3 and D4 (Supplementary Figure S1) to estimate homoeologue expression, 8,934 (86.8%) had sequences derived from both homoeologues in T2. Thus, homoeologues were retained for at least ~87% of genes initially duplicated in T2 (and almost certainly more, because some homoeologues are likely retained but not expressed highly enough under these conditions to be detected). Consequently, we estimated that T2 has 1.87 - 2.0 homoeologues per diploid gene (i.e., 1.87 - 2.0 times the number of genes per cell), but only 1.4 times the number of transcripts per cell (Figure 3.3B). Averaged across the genome, therefore, expression per gene in T2 is approximately 0.70-fold (1.4/2.0) to 0.75-fold (1.4/1.87) that of its diploid progenitors.

## **DISCUSSION**

Because transcript profiling experiments yield transcriptome-normalized expression values, they provide no information about expression per cell without knowing the relative sizes of the transcriptomes being compared. Here we have described a novel qRT-PCR assay that provides direct estimates of expression per genome and per cell, and have shown how these estimates can be coupled with transcript profiling data to obtain estimates of relative transcriptome size. These

estimates can in turn be used to determine relative expression per cell for every gene in the transcript profiling dataset.

Kanno et al. (2006), recognizing the same problem, proposed an alternative method to determine expression level per cell, but did not utilize their data to estimate relative transcriptome sizes. Also, because their focus was on normalizing microarray data, their method is necessarily less direct than ours (they used spiked RNA as a proxy for the gDNA initially present in the sample as opposed to the gDNA itself), and would require precise quantification of genome sizes before being applied to cross-species or cross-ploidy level comparisons. In contrast, the method described here is insensitive to genome size, and only requires knowledge of target gene and genome copy number per cell (ploidy level).

*Allopolyploidy and Transcriptome Size:* By coupling transcript profiling data with a genome-normalized qRT-PCR assay, we have provided the first estimates of transcriptome size (number of transcripts per cell) for several closely related species: a tetraploid and its diploid progenitors. Whereas the two diploid leaf transcriptomes are approximately the same size, that of the tetraploid is significantly larger. But despite the fact that the T2 tetraploid (*Glycine dolichocarpa*) is of fairly recent origin (within the last 100,000 years), and retains  $\geq 87\%$  of its genes in duplicate, its leaf transcriptome is only  $\sim 1.4$ -fold larger than the transcriptomes of its diploid progenitors.

It is possible that the T2 leaf transcriptome was doubled initially, and has subsequently undergone downsizing, in a process akin to genome diploidization. If so, because we observe an approximately 30% reduction in transcriptome size (vs. the sum of the two diploid transcriptomes) but only a 6% reduction in genome size (vs. the sum of the two diploid genomes), and  $\leq 7\%$  reduction in gene copy number, this suggests that transcriptome downsizing has progressed to a greater degree than

genome downsizing in this species. The transcriptome may have experienced immediate and widespread dosage compensation upon genome doubling, perhaps via epigenetic mechanisms – changes in DNA methylation have been observed in other polyploid species in the first generations following doubling (Lee and Chen 2001; Kashkush et al. 2002; Madlung et al. 2005), and chromatin modifications (histone acetylation and methylation) are associated with changes in expression of *FLC* (Wang et al. 2006b), *CCA1* and *LHY* (Ni et al. 2009) in synthetic *Arabidopsis* allotetraploids. Estimating transcriptome sizes in natural polyploids of various ages, as well as in synthetic polyploids, will shed light on this question, and reveal if changes in cellular transcript abundance are consistent across species or if they are lineage specific. Additionally, because transcriptomes vary by tissue type and growth condition, it remains to be determined whether other tissues or conditions exhibit similar responses in terms of transcriptome size.

*Dosage Responses of Individual Genes:* To date, dosage responses associated with polyploidy have only been estimated for 18 genes in a synthetic maize autopolyploid series (Guo et al. 1996). With an estimate of relative transcriptome size in hand, we were able to infer dosage responses for 15,761 genes in T2 (Figure 3.3A). In contrast to the overall pattern observed in maize (Guo et al. 1996), in which the majority of genes surveyed exhibited a 1:1 dosage effect, the majority of genes in the T2 allopolyploid (8,115; 51.5%) display an intermediate dosage response (0.8-1.3x midparent), driving the genome-wide average of partial dosage compensation. Only about 17% of the genes in T2 exhibit a 1:1 dosage response.

This difference in global dosage response pattern could be due to the hybrid origin of T2. Whereas dosage responses in maize were examined in an autopolyploid series (Guo et al. 1996), T2 was formed via interspecific hybridization, producing novel combinations of cis- and trans-acting transcriptional regulators. Alternatively,

some of the observed differences may be due to gene expression evolution in T2. Despite a relatively recent origin, T2 has been subject to natural selection for tens of thousands of years, whereas the maize polyploids were studied in the first generations following synthesis in the laboratory.

It is also possible that the limited sampling in maize (18 genes) does not provide a representative picture of overall dosage responses. Application of the methods described here to the maize synthetic autopolyploid system, as well as to other polyploidy model systems, would give a more comprehensive picture of the similarities and differences in dosage response patterns between natural and synthetic polyploids, as well as between auto- and allopolyploids.

*Modulation of Homoeologue Expression Across an Allopolyploid Genome:*

The contributions of D3 and D4 homoeologues to T2 expression could be determined for genes in which D3- or D4-specific SNPs were sequenced (Supplementary Figure S1). Thus, we were able to explore patterns of homoeologue deployment under each dosage response. In most cases both homoeologues were expressed, even when total expression was modulated to the midparent diploid level or less. Overall, one of two copies was silent for 11.5% of the homoeologue pairs examined. In a study of homoeologue expression biases in ovules of a natural cotton allotetraploid (Adams et al. 2003), only 1 of 40 pairs (2.5%) exhibited complete silencing. A more recent study using a homoeologue-specific microarray to survey the same cotton allotetraploid more broadly (Flagel et al. 2008) observed homoeologue silencing for 115 of 1,383 genes (8.3%). Thus, absolute silencing of homoeologues may be relatively rare.

Though generally uncommon, our data indicate that the frequency of homoeologue silencing varies significantly by dosage response (Figure 3.3B). The group of genes exhibiting dosage compensation (expression per cell equal to the midparent diploid expression level) had a higher frequency of homoeologue silencing

than genes exhibiting a 1:1 dosage effect (expression per cell double that of the midparent diploid expression level). Additionally, genes exhibiting extreme dosage responses in either direction ( $<0.5x$  per genome or  $>1.0x$  per genome) were significantly more likely to silence one homoeologue (21.5% and 15.9%, respectively), than genes that have undergone more moderate dosage responses (0.5x to 1.0x per genome). For genes that have experienced a negative dosage effect (expression below the diploid level per cell) this makes intuitive sense. For genes that have experienced a  $>1:1$  dosage effect, however, this result is surprising. In these cases, the polyploid is producing more than double the midparent number of transcripts per cell from the same number of loci as its diploid progenitors. Thus, complete silencing of one homoeologue is accompanied by strong up-regulation of the other.

*Relevance and Utility of Overall Transcriptome Size:* Normalizing expression data per cell provides a reliable means to compare transcript profiling experiments performed with different RNA samples, and on different platforms (Kanno et al. 2006). In addition, quantifying relative expression per cell is necessary to understand gene dosage responses, and has the potential to reveal biologically significant differences in gene regulation that may be obscured in transcriptome-normalized data.

Equivalent analyses of transcriptome size would give greater context to existing (Hegarty et al. 2005; Hegarty et al. 2006; Wang et al. 2006a; Hegarty et al. 2008) and future transcript profiling experiments comparing species and ploidy levels by making it possible to determine if additivity on a per transcriptome basis (i.e., equal transcriptome-normalized expression) translates to additivity in absolute expression. At present, different studies of gene expression in polyploids operate on the assumption that “additive” transcriptome-normalized expression represents either midparent expression (i.e., dosage compensation) or the sum of expression from the

two diploids – i.e., a 1:1 dosage effect, and often the two are used interchangeably (Jackson and Chen 2009), despite very different meanings. As our data show, either assumption could be faulty.

Recent genomic studies have led to renewed interest in gene dosage evolution (Papp et al. 2003; Blanc and Wolfe 2004; Freeling and Thomas 2006; Paterson et al. 2006; Thomas et al. 2006). Reciprocal patterns of duplicate retention following polyploidy and non-polyploid duplications suggest that dosage sensitivity is, in many cases, driving gene family evolution (Freeling 2009; Birchler and Veitia 2010). Dosage sensitivity correlates with the extent to which a gene's product forms protein-protein interactions, and the balance hypothesis correctly predicts that such “connected” genes (Thomas et al. 2006) will tend to retain polyploid duplicates and eliminate non-polyploid duplicates. There are, however, numerous exceptions. Genes that appear to meet the criteria of being connected, but do not follow the predictions of the balance hypothesis, may represent genes for which transcript abundance is readily decoupled from gene dosage (Veitia et al. 2008; Edger and Pires 2009). Consequently, cataloging dosage responses across the genome, as we have done here, will help to test and refine the balance hypothesis.

Finally, the qPCR approach utilized here, using genomic DNA to normalize expression estimates, could provide more reliable results than the typical alternative of normalizing to expression of a single reference gene in any instance where relative expression estimates are needed. Nicot et al. (2005) evaluated the stability of expression of seven housekeeping genes commonly used for RT-PCR normalization, and found significant variation in expression in response to various stresses. They concluded that only one of the seven (*Elongation factor 1- $\alpha$* ; *Elf1 $\alpha$* ) was suitable as an internal reference for the three stresses they examined. Even *Elf1 $\alpha$* , however, showed a 2-3 cycle range in threshold cycle ( $C_t$ ) between control and cold stress conditions.

Variation in the expression level of housekeeping genes has led some to recommend using combinations of genes as internal controls (Thellin et al. 1999; Vandesompele et al. 2002). This approach, however, greatly increases the size and complexity of an RT-PCR experiment.

In contrast, normalizing to gene copy number may be simpler and more reliable. Gene copy number is more stable than gene expression, and, consequently, provides a better reference for normalization. This would be true for all types of comparisons, but particularly in the case of cross-species or cross-ploidy level comparisons, where the expression levels of individual housekeeping genes might differ considerably. In a recent study of the effects of ploidy and hybridization on the circadian clock, expression estimates of central oscillator genes were normalized using *Actin2* (*ACT2*) expression (Ni et al. 2009). The possibility for variation in *ACT2* expression arising from genome doubling or hybridity was not discussed, but is potentially significant. In the present study, RNA-Seq data indicate that the combined expression of two *ACT2* orthologues in the T2 tetraploid is 1.4x the D4 diploid level on a per transcriptome basis. Thus, normalizing to *Actin* would tend to exaggerate apparent cases of down-regulation, and obscure genuine cases of up-regulation associated with polyploidy in T2. Genomic copy number is more stable than expression level (though differences in endoreduplication must be accounted for), and, arguably, more easily verified. Consequently, gene copy-normalization should provide more reliable estimates of relative expression, with the added advantage of providing direct information about dosage responses.

## **FUNDING**

This work was supported by The National Science Foundation [grant numbers IOS-0744306 and DEB-0709965 to J.J.D.].

## **ACKNOWLEDGEMENTS**

We thank Gregory D. May and Andrew D. Farmer of the National Center for Genome Resources for carrying out Illumina sequencing and primary data processing; and Daniel C. Ilut of Cornell University for writing scripts to calculate homoeologue expression levels using the RNA-Seq data. We are grateful to Brandon Gaut and three anonymous reviewers for comments on the manuscript.

## REFERENCES

- Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci* 100:4649-4654.
- Andersen MR, Vongsangnak W, Panagiotou G, Salazar MP, Lehmann L, Nielsen J. 2008. A trispecies *Aspergillus* microarray: Comparative transcriptomics of three *Aspergillus* species. *Proc Natl Acad Sci* 105:4387-4392.
- Arumuganathan K, Earle ED. 1991. Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol Biol Rep* 9:217-229.
- Barow M. 2006. Endopolyploidy in seed plants. *BioEssays* 28: 271-281.
- Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol* 186: 54-62.
- Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. *Plant Cell* 16:1679-1691.
- Blencowe BJ, Ahmad S, Lee LJ. 2009. Current-generation high-throughput sequencing: Deepening insights into mammalian transcriptomes. *Gene Dev* 23:1379-1386.
- Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol* 58:377-406.
- Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16: 738-749.
- Dolezel J, Greilhuber J, Suda J. 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* 2: 2233-2244.

- Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. 2004. Evolution of the perennial soybean polyploid complex (glycine subgenus glycine): A study of contrasts. *Biol J Linn Soc* 82:583-597.
- Edger P, Pires J. 2009. Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res* 17:699-717.
- Ehrendorfer F. 1980. Polyploidy and distribution. In: Lewis WH, editor. *Polyploidy: Biological Relevance*. NY: Plenum. p. 45-60.
- Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R et al. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* 296:340-343.
- Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc Natl Acad Sci* 106:5737-5742.
- Flagel L, Udall J, Nettleton D, Wendel J. 2008. Duplicate gene expression in allopolyploid gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol* 6:16.
- Freeling M. 2009. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* 60:433-453.
- Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* 16:805-814.
- Gilad Y, Borevitz J. 2006. Using DNA microarrays to study natural variation. *Curr Opin Genet Dev* 16:553-558.
- Gilad Y, Pritchard JK, Thornton K. 2009. Characterizing natural variation using next-generation sequencing technologies. *Trends Genet* 25:463-471.

- Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440:242-245.
- Grant V. 1981. *Plant speciation*. NY: Columbia University Press.
- Guo M, Davis D, Birchler JA. 1996. Dosage effects on gene expression in a maize ploidy series. *Genetics* 142:1349-1355.
- Hammond JP, Bowen HC, White PJ, Mills V, Pyke KA, Baker AJM, Whiting SN, May ST, Broadley MR. 2006. A comparison of the *Thlaspi caerulescens* and *Thlaspi arvense* shoot transcriptomes. *New Phytol* 170:239-260.
- Hegarty MJ, Hiscock SJ. 2008. Genomic clues to the evolutionary success of polyploid plants. *Curr Biol* 18:R435-R444.
- Hegarty MJ, Barker GL, Brennan AC, Edwards KJ, Abbott RJ, Hiscock SJ. 2008. Changes to gene expression associated with hybrid speciation in plants: Further insights from transcriptomic studies in *senecio*. *Philos T Roy Soc B* 363:3055-3069.
- Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ. 2006. Transcriptome shock after interspecific hybridization in *senecio* is ameliorated by genome duplication. *Curr Biol* 16:1652-1659.
- Hegarty MJ, Jones JM, Wilson ID, Barker GL, Coghill JA, Sanchez-Baracaldo P, Liu G, Buggs RJA, Abbott RJ, Edwards KJ et al. 2005. Development of anonymous cDNA microarrays to study changes to the *senecio* floral transcriptome during hybrid speciation. *Mol Ecol* 14:2493-2510.
- Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF. 2008a. Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. *Genetics* 179:1725-1733.
- Hovav R, Udall JA, Chaudhary B, Rapp R, Flagel L, Wendel JF. 2008b. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc Natl Acad Sci* 105:6191-6195.

- Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NWG, Couloux A, Dalwani A, Denny R et al. 2008. Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiol* 148:1740-1759.
- Jackson S, Chen ZJ. 2009. Genomic and expression plasticity of polyploidy. *Curr Opin Plant Biol* 13: 153-159.
- Kanno J, Aisaki K, Igarashi K, Nakatsu N, Ono A, Kodama Y, Nagao T. 2006. "Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays. *BMC Genomics* 7: 64.
- Kashkush K, Feldman M, Levy AA. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160:1651-1659.
- Lee H, Chen ZJ. 2001. Protein-coding genes are epigenetically regulated in arabidopsis polyploids. *Proc Natl Acad Sci* 98:6753-6758.
- Lewis WH. 1980. Polyploidy in species populations. In: Lewis WH, editor. *Polyploidy: Biological Relevance*. NY: Plenum. p. 103-144.
- Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW, Martienssen R, Comai L. 2005. Genomic changes in synthetic arabidopsis polyploids. *Plant J* 41:221-230.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509-1517.
- Miller NA, Kingsmore SF, Farmer AD, Langley RJ, Mudge J, Crow JA, Gonzalez AJ, Schilkey FD, Kim RJ, van Velkinburgh J et al. 2008. Management of high-throughput DNA sequencing projects: *Alpheus*. *J Comput Sci Syst Biol* 1:132-148.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5:621-628.

- Ni Z, Kim E, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ. 2009. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* 457:327-331.
- Nicot N, Hausman JF, Hoffmann L, Evers D. 2005. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *J Exp Bot* 56:2907-2914.
- Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. *Nat Genet* 32:261-266.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* 34:401-437.
- Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194-197.
- Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends Genet* 22: 597-602.
- Pfaffl MW, Horgan GW, Dempfle L. 2002. Relative expression software tool (REST(C)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucl Acids Res* 30:e36.
- Pfeil BE, Craven LA, Brown AHD, Murray BG, Doyle JJ. 2006. Three new species of northern Australian glycine (fabaceae, phaseolae), *G. gracei*, *G. montis-douglas* and *G. syndetika*. *Funct Plant Biol* 19:245-258.
- Quackenbush J. 2002. Microarray data normalization and transformation. *Nat Genet* 32:496-501.
- Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM. 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* 339:62-66.

- Rapp R, Udall J, Wendel J. 2009. Genomic expression dominance in allopolyploids. *BMC Biol* 7:18.
- Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the drosophila melanogaster subgroup. *Nat Genet* 33:138-144.
- Rokas A, Abbot P. 2009. Harnessing genomics for evolutionary insights. *Trends Ecol Evol* 24:192-200.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178-183.
- Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Curr Opin Plant Biol* 9:104-109.
- Stebbins GL. 1971. Chromosomal evolution in higher plants. London, UK: Edward Arnold.
- Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18:1944-1954.
- Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E. 1999. Housekeeping genes as internal standards: Use and limits. *J Biotechnol* 75:291-295.
- Thomas BC, Pedersen B, 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.
- Udall JA, Swanson JM, Nettleton D, Percifield RJ, Wendel JF. 2006. A novel approach for characterizing expression levels of genes duplicated by polyploidy. *Genetics* 173:1823-1827.

- Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3: research0034.1-research0034.11.
- Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: Genomic, transcriptomic and proteomic effects. *Trends Genet* 24:390-397.
- Wang J, Tian L, Lee HS, Wei NE, Jiang H, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L et al. 2006a. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics* 172:507-517.
- Wang J, Tian L, Lee H, Chen ZJ. 2006b. Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in arabidopsis allopolyploids. *Genetics* 173:965-974.
- Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploidy speciation in plants. *Proc Natl Acad Sci* 106: 13875-13879.
- Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873-881.

## CHAPTER 4

### **Enhanced photoprotection in a recent allotetraploid correlates with high levels of galactolipid synthase gene expression and modified galactolipid profiles**

#### **ABSTRACT**

Allopolyploidy increases photosynthetic capacity across a range of plant taxa, and is also believed to confer enhanced stress tolerance. Excess light is a ubiquitous plant stress that accompanies photosynthetic light reactions. We quantified the photoprotective capacities ( $NPQ_{max}$ ) of a recently formed natural allotetraploid (*Glycine dolichocarpa*) and its diploid progenitors (*G. tomentella* and *G. syndetika*), and showed that the tetraploid has a greater capacity for NPQ when grown under chronic excess light. We further showed that this increase in  $NPQ_{max}$  was due to an increase in protective photoinhibition ( $qI_P$ ). Using transcript profiling data from the same species and treatments, we excluded most known NPQ genes as driving the enhanced NPQ phenotype, but showed that two gene families involved in galactolipid biosynthesis (*MGD* and *DGD*) exhibited expression patterns consistent with a role in increasing NPQ. In accordance with these expression profiles, lipid profiling indicated that the tetraploid has higher steady state levels of digalactosyldiacylglycerol (DGDG), as well as a greater relative increase in monogalactosyldiacylglycerol (MGDG) in response to excess light than either diploid progenitor.

#### **INTRODUCTION**

At least 30% of flowering plants are polyploid (Soltis et al. 2009), and an estimated 15% of all angiosperm speciation events involved polyploidy (Wood et al.

2009). Polyploids frequently exhibit greater colonizing ability and expanded geographical ranges compared to their diploid progenitors (Doyle et al. 2004, Lewis 1980, Otto and Whitton 2000). One explanation for the prevalence and apparent success of polyploids in flowering plants is that they possess greater stress tolerances than their diploid progenitors, making them better adapted to extreme and adverse environments (Ehrendorder 1980, Lewis 1980, Otto and Whitton 2000).

Plants routinely absorb more energy from sunlight than they can use for photosynthesis (Demmig-Adams and Adams, 2000; Avenson et al., 2004). Excess light energy can quickly generate reactive oxygen species (ROS) that damage the pigments, proteins and lipids of the photosynthetic apparatus (Niyogi, 2000). Thus, excess light stress is a ubiquitous plant stress that is an inevitable byproduct of photosynthetic light harvesting.

Photosynthesis as a whole is known to be affected by polyploidy (reviewed by Warner and Edwards, 1993). Studies involving a wide range of polyploid species have shown that polyploids often have larger mesophyll cells with more chloroplasts, higher chlorophyll and rubisco contents, and greater photosynthetic capacities per cell, than their diploid counterparts.

One aspect of photosynthesis that has not been studied in the context of polyploidy is photoprotection, or the mechanisms by which plants avoid damage from excess absorbed light energy. These mechanisms include avoiding excess light via leaf and/or chloroplast movement, preventing the formation of ROS by thermal dissipation of excess light energy, and ROS scavenging to prevent photooxidative damage (Li and Niyogi 2009). Because thermal dissipation of excess light energy can be measured as quenching of chlorophyll fluorescence, it is also known as nonphotochemical quenching (NPQ). Several strategies for NPQ are evolutionarily conserved throughout the plant kingdom (Avenson et al., 2004; Horton and Ruban, 2005), and are essential

for plant fitness in the field (e.g., Kulheim et al., 2002).

NPQ pathways are centered in photosystem II (PSII), and include a rapidly reversible component termed energy-dependent quenching (qE), and slowly reversible or irreversible components collectively termed photoinhibition (qI) (Niyogi 2000). Photoinhibition consists of a protective component (qI<sub>p</sub>) that is reversible within a few hours in the dark, and a damage component that is only reversed by the PSII damage-repair cycle (Melkonian et al. 2004). Because this cycle requires light-dependent protein synthesis, qI<sub>D</sub> is irreversible in the dark (Melkonian et al. 2004).

The non-damage components of NPQ are not fully understood, but are believed to involve conformational changes in the PSII light harvesting complex (LHCII) (Horton et al. 2008, Szabo et al. 2005). qE is known to require the PSII subunit, PsbS (Li et al. 2000, 2002), as well as the xanthophyll cycle enzymes, violaxanthin deepoxidase (VDE), zeaxanthin epoxidase (ZE), and lycopene cyclase (LYCE) for optimal function (Szabo et al. 2005).

PSII is embedded in the thylakoid membrane, and is intimately associated with the major thylakoid lipids, monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG) (Benning and Ohta 2005, Loll et al. 2005, Holzl et al. 2009). These lipids are synthesized by MGDG synthase and DGDG synthase, respectively. MGDG catalyzes the galactosylation of diacylglycerol (DAG) to produce MGDG, and MGDG is in turn the substrate for galactosylation by DGDG to produce DGDG (Benning and Ohta 2005). Both MGDG and DGDG have recently been implicated in NPQ (Aronsson et al. 2008, Holzl et al. 2009).

Adaptations that increased the capacity of the photoprotective apparatus or its flexibility to adjust to different light environments could potentially have contributed to the success and expanded ranges of some polyploids. Here we examined the effects of allopolyploidy on NPQ by quantifying NPQ capacity, gene expression of several

know NPQ genes, and leaf lipid profiles under limiting and excess light in a recently formed allotetraploid (*Glycine dolichocarpa*) and its diploid progenitors (*G. tomentella* and *G. syndetika*) (Doyle et al. 2004).

## METHODS

*Plant material:* The study group consisted of the natural allopolyploid, *Glycine dolichocarpa* ( $2n = 80$ ; designated “T2”) and its diploid progenitors, *G. tomentella* ( $2n = 40$ ; “D3”) and *G. syndetika* ( $2n = 40$ ; “D4”). (Doyle et al. 2004; Pfeil et al. 2006). The two diploid species, D3 and D4, diverged approximately 2.5 MYA, and hybridized to give rise to T2 within the last 100,000 years (Doyle et al. 2004).

Plants were grown in common growth chambers with a 12hr/12hr light/dark cycle, 22°C/18°C day/night temperature regime and a light intensity of either 125  $\mu\text{mol m}^{-2} \text{s}^{-1}$  (limiting light; LL) or 800  $\mu\text{mol m}^{-2} \text{s}^{-1}$  (excess light; EL). 800  $\mu\text{mol m}^{-2} \text{s}^{-1}$  was chosen as the intensity for EL based on measurements of linear electron transport (ETR) that showed that all genotypes approach light saturation at 800  $\mu\text{mol m}^{-2} \text{s}^{-1}$  or less (data not shown). All leaves sampled for chlorophyll fluorescence, transcript profiling and lipid profiling were all of an equivalent developmental stage (young but fully expanded).

*Chlorophyll fluorescence:* We measured NPQ in the allopolyploid (T2) and its diploid progenitors (D3 and D4) using a Hansatech FMS2 chlorophyll fluorometer. Leaves were dark adapted for  $\geq 15$  minutes prior to measurement in order to obtain an estimate of maximal fluorescence ( $F_m$ ) and maximum quantum yield of photosystem II ( $F_v/F_m$ ). If a leaf was photoinhibited ( $F_v/F_m < 0.8$ ), it was placed under low light (approx. 30  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ) for 1-2 hours to allow for  $q_{I_p}$  relaxation and PSII repair prior to measuring NPQ. Leaves that were still photoinhibited after this recovery period were discarded.

Maximum NPQ ( $\text{NPQ}_{\text{max}}$ ) was determined by measuring NPQ under non-photosynthetic conditions (a gas mixture of 0%  $\text{CO}_2$ , 2%  $\text{O}_2$  to allow build up of the trans-thylakoid pH gradient necessary for qE, and 98%  $\text{N}_2$ ) and high light ( $2000 \mu\text{mol m}^{-2}\text{s}^{-1}$ ) (Demmig-Adams et al., 2006). Measurements were performed on detached leaves in a sealed chamber to allow regulation of the gas mixture.  $F_m'$  was measured under these conditions every 15-60 seconds for 12 minutes.

NPQ was calculated as  $(F_m - F_m')/F_m'$  (Maxwell and Johnson 2000). Curves were fit to the data (NPQ vs. time) for each leaf by non-linear regression using SigmaPlot. All data were fit to a two-phase model of exponential rise to a maximum:

$$\text{NPQ} = a (1 - e^{-b*x}) + c (1 - e^{-d*x}).$$

This model was chosen because NPQ induction is known to involve two distinct phases, the first triggered by protonation of the PsbS protein, and the second resulting from deepoxidation of violaxanthin to zeaxanthin (Horton et al. 2008). Correlation coefficients ( $r^2$ ) were consistently higher for this model than for single component models, and residuals were more randomly distributed.

The relative contributions of energy-dependent quenching (qE), and photoinhibitory quenching (qI) to  $\text{NPQ}_{\text{max}}$  were determined by monitoring NPQ relaxation kinetics in the dark following treatment, as described (Melkonian et al. 2004). Briefly, following the 12 minute exposure to saturating light, leaves were left in darkness, and  $F_m'$  re-measured after 10 minutes and again after 2-4 hours. Because qE relaxes within 10 minutes of being placed in darkness, NPQ persisting after 10 minutes in the dark consists of qI ( $qI_P + qI_D$ ), and the difference between  $\text{NPQ}_{\text{max}}$  and NPQ after 10 minutes is equal to qE. Because the protective component of photoinhibition ( $qI_P$ ) relaxes within two hours of being placed in darkness, NPQ persisting after  $\geq 2$  hours in the dark consists of the damage component of

photoinhibition ( $qI_D$ ), which does not relax in darkness due to the requirement for light-dependent protein synthesis.

Consequently,  $qI_{total}$  was calculated as  $(F_m - F_{m1}') / F_{m1}'$ , where  $F_{m1}'$  is the value of  $F_m'$  measured after 10 minutes in the dark following measurement of  $NPQ_{max}$ , and  $qI_D$  was calculated as  $(F_m - F_{m2}') / F_{m2}'$ , where  $F_{m2}'$  is the value of  $F_m'$  measured after  $\geq 2$  hours in the dark following measurement of  $NPQ_{max}$ .  $qE$  was calculated as  $NPQ_{max} - qI_{total}$ , and  $qI_P$  was calculated as  $qI_{total} - qI_D$ .

At least three measurements were made per plant (three separate leaves), and three plants were measured per species. The order in which plants were measured was randomized to reduce potentially confounding effects of diurnal variation in photosynthesis.

*Transcript profiling:* Young, fully expanded leaflets were collected 1.5 – 2.0 hours into the light period and immediately frozen in liquid nitrogen. Leaflets were pooled from six individuals per species, and RNA was isolated using the Qiagen Plant RNeasy kit with on-column DNase treatment. Sequencing was performed using Solexa/Illumina “Sequencing by Synthesis” at the National Center for Genome Resources with the following modifications. Poly A+ RNA was annealed to high concentrations of random hexamers, reverse transcribed, and ligated to adapters complementary to sequencing primers. The cDNA was then amplified by 20 cycles of polymerase chain reaction and size fractionated on agarose gels. 200 bp amplicons were excised and sequenced by synthesis with reversible terminator nucleotides with cleavable fluorescence.

To process the data for analysis, files were mirrored to an off-instrument computer using the Illumina® platform to perform image analysis, base-calling, quality filtering, and per base confidence scores. Sequences were then aligned using GSNAP (Wu and Nacu 2010) against the 8X genome sequence of soybean (*Glycine*

*max*) (Schmutz et al. 2010), which diverged from the common ancestor of D3, D4 and T2 approximately 5 MYA (Innes et al. 2008). Note that soybean, D3, and D4, all of which are  $2n = 40$ , are fully diploidized descendants of an ancestor that underwent a whole genome duplication  $\leq 13$ MYA (Schmutz et al. 2010). Roughly half of the genes duplicated by this event are retained in duplicate in the soybean genome (Schmutz et al. 2010). Only reads mapping unambiguously to a single copy in the soybean genome were used in this study.

GSNAP was parameterized to allow spliced alignments of the transcript reads to the genomic reference sequences requiring canonical splice sites and allowing introns of up to 10Kbp; alignments were also allowed to include small indels and mismatches, but required that at least 30 out of the 36 base pairs in a read were matched. Alignments above this threshold with the highest number of identities were divided into three classes: uniquely aligned reads, low-copy repetitive alignments matching no more than 5 locations in the reference and highly repetitive reads matching  $>5$  locations in the reference. The alignments in the first two classes were further processed using the Alpheus pipeline (Miller et al. 2008) for deriving per-gene read counts and sequence polymorphism calls. The boundaries of each gene were taken as the maximal starting and ending positions from any of the transcripts associated with the gene, and any read alignment partially contained within this span was counted toward the expression of that gene in the given sample. Reads from uniquely aligned sequences were used to estimate expression levels after normalizing read counts to account for overall sampling sizes. Transcript abundance per transcriptome for a given gene was estimated as the number of reads unambiguously mapped to that gene per million unambiguously mapped reads generated by that library (reads per million; RPM). Differences in expression were tested for

significance using the method of Audic and Claverie (1997) with a Bonferroni correction for multiple comparisons.

*qRT-PCR*: For selected genes, relative expression estimates from the RNA-Seq experiment were validated by qRT-PCR. As with the RNA-Seq experiments, young, fully expanded leaflets were collected 1.5 – 2.0 hours into the light period and immediately frozen in liquid nitrogen. RNA was isolated using the Qiagen Plant RNeasy kit with on-column DNase treatment, and reverse transcribed with random decamers using the Ambion Retroscript kit. The cDNA was diluted five-fold and used as template for qRT-PCR with the following components: 5.75ul H<sub>2</sub>O, 7.5ul Power SYBR Green master mix (Applied Biosystems), 0.375ul forward primer, 0.375ul reverse primer, and 1ul template. Assays were performed on an Applied Biosystems 7900 HT instrument, with 40 PCR cycles. Dissociation curves were generated at the end of the PCR to confirm specificity of amplification. For each primer pair and species, we amplified three technical replicates from each of three biological replicates.

Amplification efficiencies were estimated using LinRegPCR (Ramakers et al. 2003) for each individual reaction. Mean efficiency per amplicon was used for relative expression estimates. For cross-species comparisons, expression of each target gene (cDNA-specific amplification) was normalized to actin expression. Relative expression values (T2/D3, T2/D4, T2/midparent, and D4/D3) were estimated using the Relative Expression Software Tool (REST) (Pfaffl et al. 2002).

*Lipid profiling*: Young, fully expanded leaflets were collected 1.5 – 2.0 hours into the light period and immediately frozen in liquid nitrogen. Two separate pools of leaflets from 3-4 individuals were collected per species. Tissue was lyophilized for 24 hours, and then shipped on dry ice to the laboratory of E. Marechal at CEA-Grenoble

(France) for lipid profiling. Lipid profiles were characterized as described (Awai et al. 2001).

## RESULTS

*NPQ capacity:* The allotetraploid (T2) and its diploid progenitors (D3 and D4) exhibited comparable NPQ capacities ( $NPQ_{max}$ ) when grown under limiting light (LL) ( $p = 0.844$ ; ANOVA) (Figure 4.1A). When grown under excess light (EL), all three species responded by increasing  $NPQ_{max}$  (Figure 4.1B). However, the increase was doubled in T2 relative to the diploids ( $\sim 33\%$  vs.  $\sim 16\%$ ) (Figure 4.1B), and  $NPQ_{max}$  under EL was significantly higher in T2 than in either diploid ( $p \leq 0.006$ ; Bonferroni T-test).

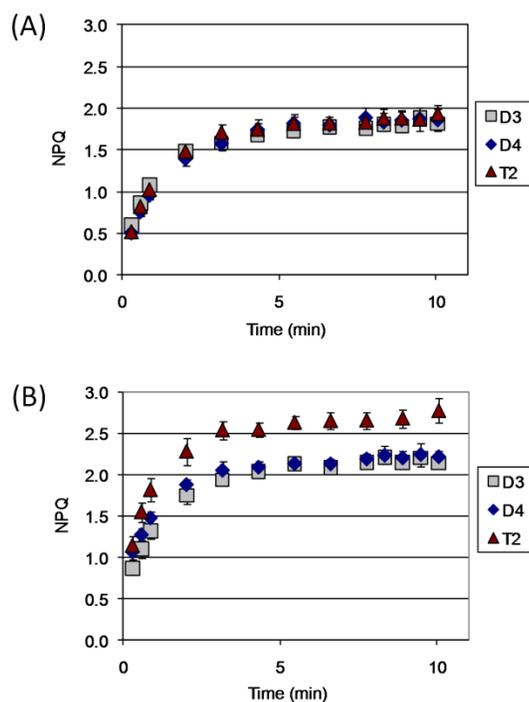


Figure 4.1. NPQ response curves ( $\pm$  SE;  $N=3$ ) for D3, D4 and T2 grown under limiting light (A) or excess light (B).

$qE$ ,  $qI_P$  and  $qI_D$ : By monitoring the relaxation kinetics of NPQ in the dark, we dissected  $NPQ_{max}$  into its component parts: energy-dependent quenching ( $qE$ ), protective photoinhibition ( $qI_P$ ) and damage-induced photoinhibition ( $qI_D$ ). Under LL, we observed no differences between D3, D4 and T2 in the contributions of these individual components to  $NPQ_{max}$  ( $p \geq 0.26$ ; ANOVA) (Figure 4.2). As has been observed in other species (Niyogi et al., 2005; Szabo et al., 2005),  $qE$  was the major component (73-82% ) of NPQ in D3, D4 and T2. The damage component of photoinhibition ( $qI_D$ ) comprised the next largest portion (13-19%), and  $qI_P$  was the smallest portion (4-5%) of total NPQ in all three species (Figure 4.2).

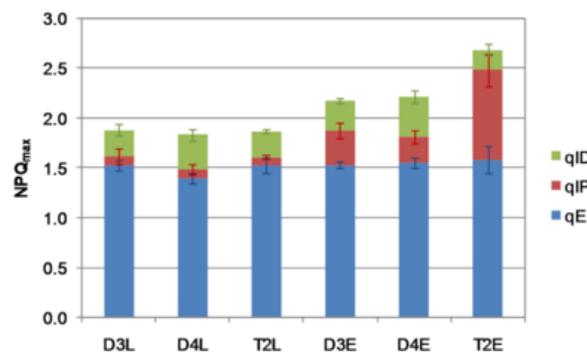


Figure 4.2. Contributions of energy-dependent quenching ( $qE$ ), protective photoinhibition ( $qI_P$ ) and damage-induced photoinhibition ( $qI_D$ ) to total NPQ ( $\pm SE$ ;  $N=3$ ). L = limiting light; E = excess light.

Under EL,  $qE$  remained the major component of total NPQ in all three species, though the total amount of  $qE$  remained constant, and there were no significant differences among the three species ( $p = 0.899$ ; ANOVA). Similarly, no significant differences were detected in  $qI_D$  under EL ( $p = 0.100$ ; ANOVA), though it is noteworthy that  $qI_D$  increased in both D3 and D4 in EL vs. LL, but decreased in T2 (Figure 4.2). In all three species,  $qI_P$  was greater in plants grown under EL than in

plants grown under LL. The increase was greatest in T2, and total  $qI_P$  was significantly greater in T2 than in D3 or D4 under EL ( $p \leq 0.035$ ; Bonferroni T-test) (Figure 4.2). Thus, the significant increase in  $NPQ_{max}$  in T2 relative to D3 or D4 under EL was due almost entirely to enhanced  $qI_P$  (Figure 4.2). As a consequence, the fraction of  $NPQ_{max}$  contributed by the different components differed between T2 and its progenitors in EL. The fraction of total NPQ contributed by  $qI_P$  was significantly higher in T2 than in D4 ( $p = 0.036$ ; Bonferroni T-test) and nearly significantly higher in T2 than in D3 ( $p = 0.088$ ; Bonferonni T-Test). The fraction of total NPQ resulting from  $qI_D$  was significantly lower in T2 than D4 ( $p = 0.024$ ; Bonferonni T-test), and numerically but not significantly lower than D3 ( $p = 0.170$ ; Bonferonni T-test). Similarly, whereas total  $qE$  was comparable among the three species, the fraction of total NPQ contributed by  $qE$  was lower in T2 than in D3 or D4, though the difference did not quite pass the significance threshold ( $p = 0.051$ ; ANOVA) (Figure 4.2).

*Transcript profiling:* In order to begin identifying the genetic basis for observed differences in NPQ, we profiled the transcriptomes of T2, D3 and D4 under LL and EL using RNA-Seq. For each of the six species/treatment combinations, RNA was extracted from leaf tissue pooled from six individuals, including the three individuals used for NPQ measurements.

High throughput sequencing using Solexa/Illumina technology generated  $> 5$  million 36bp reads for each sample (species/treatment combination). Reads were uniquely mapped to  $> 35,000$  genes in each sample, with unique read counts per gene ranging from 1 to  $> 98,000$ , reflecting the relative abundance of that transcript in the transcriptome (Marioni et al. 2008). The expression level per transcriptome for a given gene was estimated as the number of sequencing reads derived from that gene divided by the total number of reads derived from that sample, reported as reads per million (RPM).

To assess the accuracy of the RNA-Seq expression data, we measured within-sample estimates of relative expression for five genes by qRT-PCR and compared these to the RNA-seq based estimates (Figure 4.3). In this case, because we compared expression levels across genes within a sample, RNA-Seq expression values were normalized by gene length to give reads per kilobase per million reads (RPKM) (Mortazavi et al. 2008). In all six libraries (LL and EL libraries for T2, D3 and D4), strong correlations were observed between RNA-Seq and qRT-PCR ( $r^2 > 0.96$ ), indicating that RNA-Seq expression estimates were accurate. For all subsequent analyses, we compared the relative expression of individual genes across species (as opposed to comparing across multiple genes within a species). Consequently, relative expression estimates were not affected by variation in gene length, making length adjustments (e.g., RPKM) unnecessary, and RPM were used.

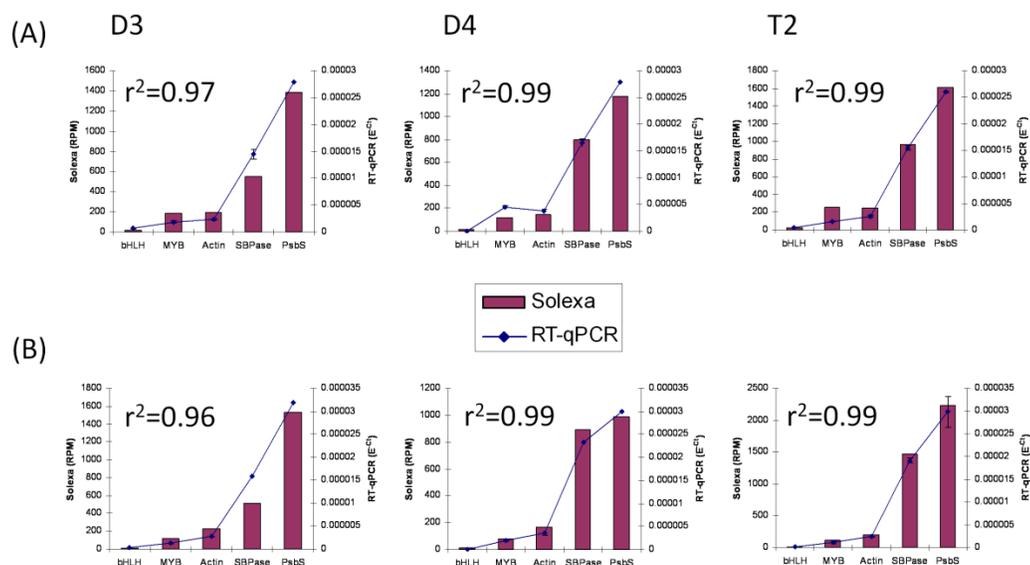


Figure 4.3. qRT-PCR validation of RNA-Seq expression estimates for five genes. (A) Limiting light; (B) Excess light.

Using the RNA-Seq data, we examined the expression profiles of several genes that are known to play important roles in NPQ. In *Arabidopsis*, the PsbS protein is required for qE (Li et al., 2000), and is a key determinant of fitness under excess light (Kulheim et al., 2002; Li et al., 2002). In the two diploids, the cumulative expression of *PsbS* decreased subtly under EL vs. LL (EL expression is 0.8 – 0.9-fold LL expression), whereas in T2 *PsbS* expression increased subtly (1.2-fold). Cumulative *PsbS* expression in EL was 1.1-fold greater in T2 than D3 and 1.8-fold greater than D4 (Figure 4.4). These differences, though small, were statistically significant ( $p < 0.05$ ).

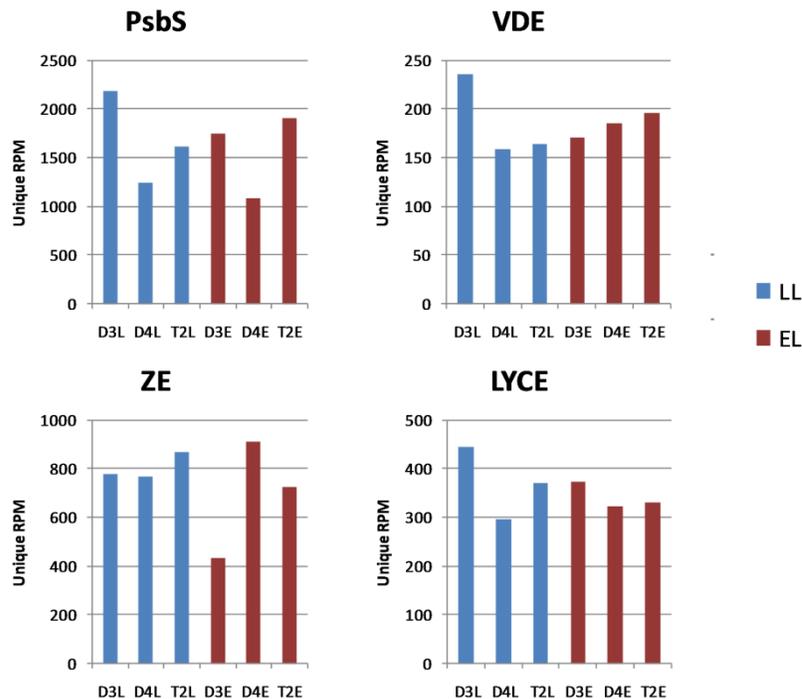


Figure 4.4. RNA-Seq based estimates of combined expression for gene families encoding *PsbS* and xanthophyll cycle enzymes.

Mutants defective in the xanthophyll cycle proteins, violaxanthin deepoxidase (VDE), zeaxanthin epoxidase (ZE), and lycopene cyclase (LYCE) are also impaired in NPQ (Anwaruzzaman et al., 2004). Under EL, minimal expression differences were

observed between T2 and the diploids for *VDE* or *LYCE* ( $p > 0.05$ ; Bonferroni) (Figure 4.4). *ZE* expression was 1.7-fold higher in T2 than D3 under EL ( $p < 0.05$ ), but 0.8-fold lower than D4 ( $p < 0.05$ ). *ZE* expression decreased in EL relative LL in D3 and T2 (Figure 4.4).

Mutants in some PSII light harvesting proteins (LhcB) also exhibit impairment of NPQ in *Arabidopsis* (Horton et al. 2008). With one exception, T2 exhibited expression comparable to either diploid under EL for each of these gene families (0.8 to 1.1-fold) (Figure 4.5). The exception was *LhcB3*, for which T2 expression was 2.1-fold greater than D4 under EL ( $p < 0.05$ ). However, *LhcB3* expression was also significantly higher in T2 than D4 under LL (3.2-fold;  $p < 0.05$ ), suggesting that the difference in *LhcB3* expression under EL does not explain the difference in NPQ response to EL between these species.

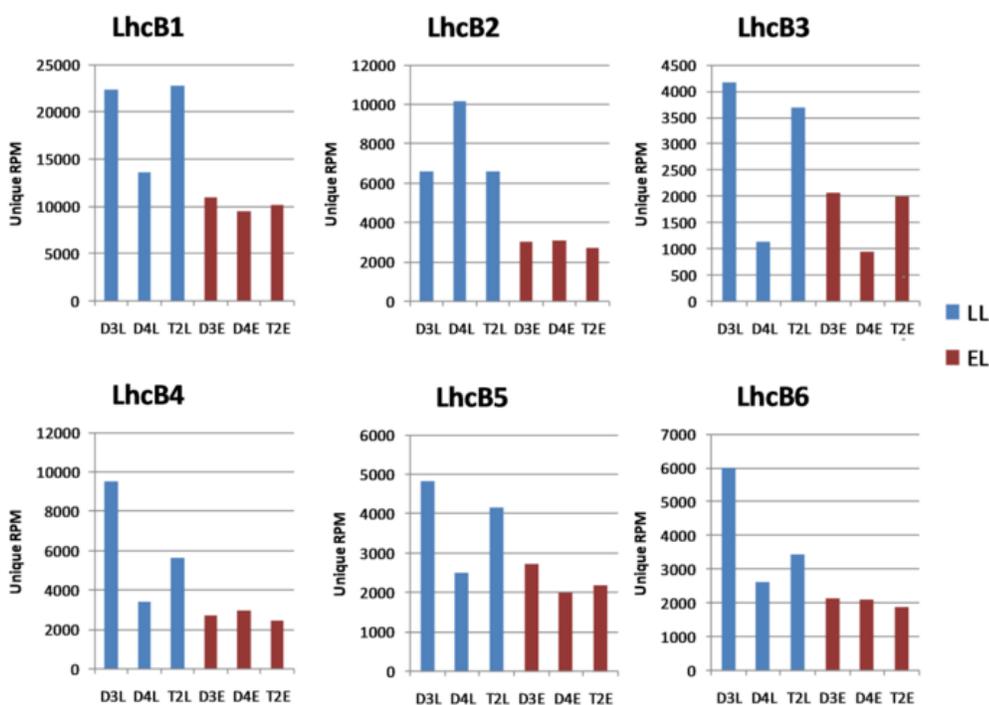


Figure 4.5. RNA-Seq based estimates of combined expression for gene families encoding photosystem II-associated light harvesting proteins (LhcB1 - LhcB6).

Because NPQ takes place in photosystem II (PSII) we then examined the expression patterns of genes encoding all of the nuclear-encoded subunits of PSII. Under EL, minimal differences in expression were observed between T2 and D3 across the PSII subunits (0.9 – 1.3-fold; Figure 4.6). T2 expression deviated to a greater extent from D4 (Figure 4.6). *PsbY* expression was more than two-fold greater in T2 than D4 ( $p < 0.05$ ). Unlike *LhcB3*, no difference was observed between T2 and D4 in *PsbY* expression under LL. Total *PsbY* expression decreased significantly under EL in both diploids, but not in T2.

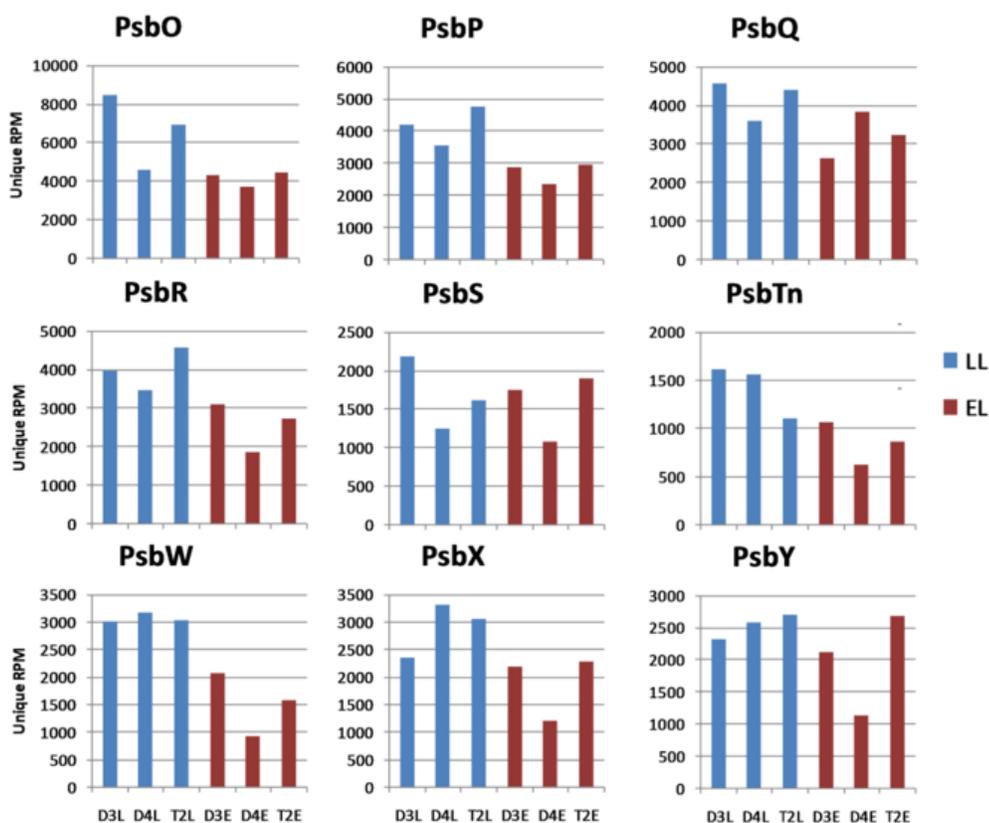


Figure 4.6. RNA-Seq based estimates of combined expression for gene families encoding photosystem II (PSII) subunits (PsbO - PsbY).

NPQ was also recently shown to be impaired in *Arabidopsis* mutants defective in biosynthesis of the major thylakoid lipids, MGDG (Aronsson et al. 2008) and DGDG (Holzl et al. 2009). In *Arabidopsis*, MGDG synthesis is catalyzed by the MGDG synthase (MGD) gene family, which is divided into A-type and B-type enzymes, with the A-type genes encoding the major isoforms in green tissue (Benning and Ohta 2005). There are two copies of each type in *Glycine* resulting from a whole genome duplication  $\leq 13$ MYA (Schmutz et al. 2010). Under LL, A-type gene expression was greater than B-type expression in all three species (Figure 4.7). Combined expression of the A-type genes was slightly higher in D3 than in D4 (1.6-fold;  $p = 0.001$ ) or T2 (1.4-fold;  $p = 0.007$ ) (Figure 4.7A). The A-type genes exhibited no light response, and D3 expression remained marginally higher than D4 (1.2-fold;  $p = 0.001$ ) or T2 (1.2-fold;  $p = 0.085$ ). RNA-Seq based expression estimates were confirmed by qRT-PCR with three biological replicates (Figure 4.7B).

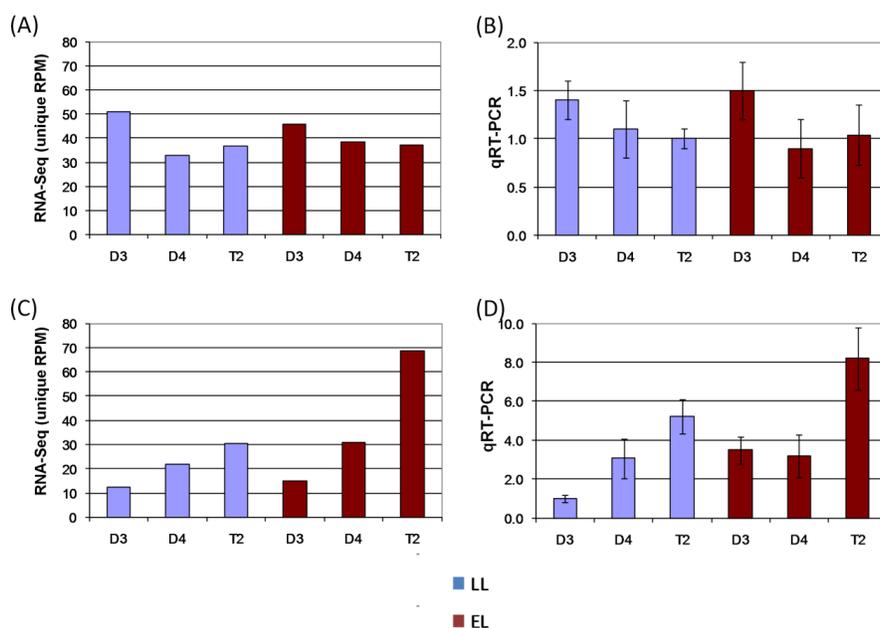


Figure 4.7. RNA-Seq and qRT-PCR estimates of MGD gene expression. (A) Type A, RNA-Seq; (B) Type A, qRT-PCR; (C) Type B, RNA-Seq; (D) Type B, qRT-PCR.

In contrast, expression of both B-type genes individually, as well as combined B-type expression, was significantly higher in T2 than in D3 or D4 under EL (1.9 to 5.4-fold higher;  $p < 0.001$ ) (Figure 4.7C). In addition, combined B-type expression was significantly up-regulated in T2 in response to EL (2.3-fold;  $p < 0.001$ ), but showed a weaker light response in D3 (1.2-fold;  $p = 0.05$ ) and D4 (1.4-fold;  $p = 0.006$ ). Consequently, most MGD expression was derived from the A-type genes in the diploids under both LL and EL, but the B-type genes were more highly expressed in T2 under EL. RNA-Seq based expression patterns for B-type MGD were confirmed by qRT-PCR (Figure 4.7D).

Synthesis of DGDG from MGDG is catalyzed by DGDG synthase (DGD). There are 10 DGD-like genes in *Glycine* (Schmutz et al. 2010), but > 80% of DGD expression was derived from two genes (Glyma03g36050 and Glyma19g38720) in all three species. Under LL, combined expression of these genes was significantly up-regulated in T2 relative to D3 (1.5-fold;  $p = 0.001$ ), and marginally up-regulated relative to D4 (1.1-fold;  $p = 0.045$ ) (Figure 4.8). Under EL, combined expression of the *DGD* genes was significantly up-regulated in T2 relative to both D3 and D4 (2.0 to 2.1-fold;  $p = 0.001$ ).

In addition, combined *DGD* expression was significantly up-regulated in T2 in response to EL (1.7-fold;  $p = 0.001$ ), but showed a weaker light response in D3 (1.3-fold;  $p = 0.015$ ) and exhibited no light response in D4 (0.9-fold;  $p = 0.075$ ) (Figure 4.8). Up-regulation of *DGD* in T2 relative to D3 was confirmed by semi-quantitative RT-PCR (Figure 4.8) and qRT-PCR (not shown). No *DGD* amplification was detected in D4 (Figure 4.8), most likely indicating that the *DGD* PCR primers did not anneal in this species.

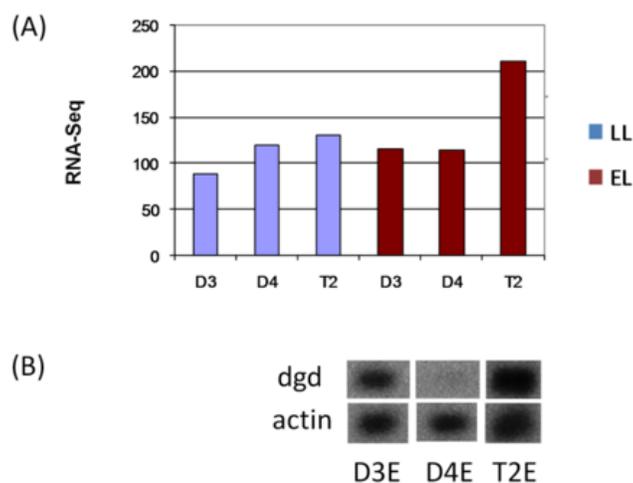


Figure 4.8. RNA-Seq and semi-quantitative RT-PCR estimates of combined DGD gene expression.

*Lipid profiling:* Because galactolipids have been shown to play a role in NPQ, and our expression data indicate that galactolipid synthase genes are more highly expressed in T2 than D3 or D4 under EL, we examined the leaf lipid profiles of each species. MGDG comprised a significantly greater fraction of total plastid lipid content in D3 than in D4 or T2 under both LL and EL (Figure 4.9A). This was consistent with D3 having slightly higher expression of A-type *MGD* genes than D4 or T2, but is somewhat surprising given that T2 exhibits a higher total level of *MGD* expression (combining A-type and B-type expression).

In all three species, the MGDG fraction of plastid lipids increased under EL, consistent with the putative photoprotective function of MGDG (Aronnson et al. 2008). T2 exhibited the largest relative increase (23.2% compared to 13.9% for D3 and 13.8% for D4) (Figure 4.9A).

Consistent with the *DGD* expression profiles, DGDG comprised a significantly greater fraction of total plastid lipid in T2 than in D3 or D4 under LL ( $p \leq 0.039$ ; Bonferonni) (Figure 4.9B). DGDG also comprised a significantly greater fraction of

plastid lipid in T2 than in D3 under EL ( $p = 0.029$ ; Bonferroni), and a slightly but not significantly higher fraction than in D4 (Figure 4.9B).

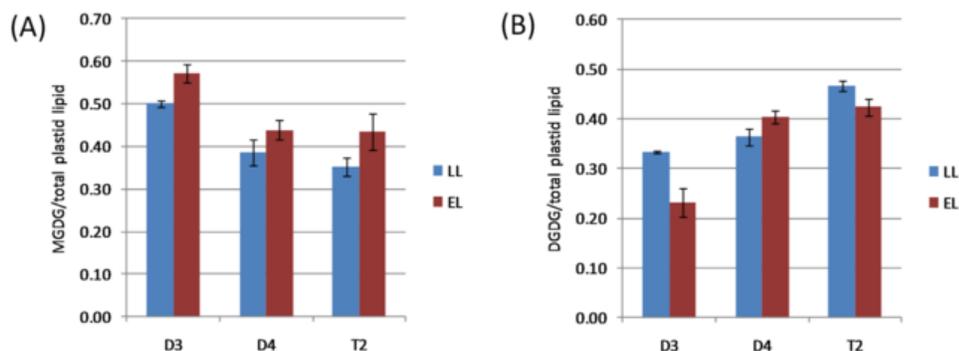


Figure 4.9. Leaf galactolipid profiles under LL and EL. (A) MGDG content as a fraction of total plastid lipid. (B) DGDG content as a fraction of total plastid lipid.

## DISCUSSION

Polyploidy has been shown to increase photosynthetic capacity in many taxa (citations in Warner and Edwards 1993), but no previous studies have examined the effects of polyploidy on the capacity to manage excess light stress, an unavoidable consequence of photosynthesis. Here we showed that under conditions of chronic excess light, photoprotective capacity ( $NPQ_{max}$ ) is significantly enhanced in a recently formed natural allotetraploid, *G. dolichocarpa* (T2) relative to its diploid progenitors, *G. tomentella* (D3) and *G. syndetika* (D4). When grown under light conditions that are limiting to photosynthesis, and therefore unlikely to result in significant excess light stress, the polypliod exhibited  $NPQ$  capacity comparable to its diploid progenitors. Therefore, the observed increase in  $NPQ_{max}$  represents a capacity to acclimate to excess light rather than a constitutively activated trait that would be maladaptive under limiting light conditions. Polypliods frequently exhibit broader geographic ranges than their diploid progenitors, leading some to speculate that they

have a greater ability to tolerate stress (Otto and Whitton 2000). In this case, T2 has a range that extends beyond Australia, to which its diploid progenitors (D3 and D4) are confined. Our data provide a concrete example of how a polyploid is better equipped to tolerate stress than its diploid progenitors.

By monitoring the kinetics of NPQ relaxation, we were further able to dissect  $NPQ_{max}$  into its component parts. Energy-dependent quenching (qE) is typically the major component of NPQ (Niyogi et al., 2005; Szabo et al., 2005), and constituted 73-82% of total NPQ in each of the species examined here. However, despite an increase in  $NPQ_{max}$  under EL, qE remained constant in all three species. This is perhaps not surprising as qE is a rapidly reversible form of NPQ that is thought to be of greatest importance in dealing with rapidly fluctuating light conditions (Kulheim et al. 2002). Under the chronic excess light conditions in which these plants were grown, it would make sense that the overall strategy for energy dissipation would shift to a more sustained mechanism.

Protective photoinhibition (qI<sub>P</sub>) is a more slowly relaxing form of NPQ than qE, yet is still reversible (i.e., does not represent damage to the PSII reaction center) (Melkonian et al. 2004). In all three species, qI<sub>P</sub> was greater in plants grown under EL than in plants grown under LL. The increase was greatest in T2, and total qI<sub>P</sub> was significantly greater in T2 than in D3 or D4 under EL (Figure 4.2). Thus, the significant increase in  $NPQ_{max}$  in T2 relative to D3 or D4 under EL was due almost entirely to enhanced qI<sub>P</sub> (Figure 4.2). Though we failed to detect significant differences in the damage component of photoinhibition (qI<sub>D</sub>), this component increased in both D3 and D4 in EL vs. LL, but decreased in T2 (Figure 4.2). Consequently, the enhanced capacity for NPQ in T2 was accompanied by less damage to the PSII reaction center.

Using transcript profiling data for T2, D3 and D4 under the two light conditions examined here, we examined the expression of several genes known to play roles in NPQ in other species. In the two diploids, the cumulative expression of *PsbS* decreased subtly under EL vs. LL, whereas in T2 *PsbS* expression increased subtly, and *PsbS* expression in EL was only modestly higher in T2 than in either diploid. It is possible that these moderate differences in *PsbS* expression explain the observed differences in NPQ phenotype. However, despite greater total NPQ in T2 under EL, the amount of qE was unaltered. Because previous studies have shown that qE is limited by PsbS protein level (Li et al. 2000), and qE was doubled in an *Arabidopsis* mutant over-expressing *PsbS* (Li et al. 2002), this suggests that the observed slight increase in *PsbS* expression in T2 was not biologically significant.

Similarly, with few exceptions, generally minimal differences were observed in the expression of genes encoding xanthophyll cycle enzymes (VDE, ZE, LYCE), light harvesting proteins (*Lhcs*), or other subunits of PSII. Under EL, *ZE* expression was 1.7-fold higher in T2 than in D3. However, total *ZE* expression decreased from LL to EL in both D3 and T2, making it unlikely that *ZE* expression patterns explain the observed increase in NPQ under EL. Similarly, *LhcB3* expression was 2.1-fold greater in T2 than in D4 under EL, but *LhcB3* expression decreased in all three species going from LL to EL. In addition, the difference in *LhcB3* expression between T2 and D4 was greater under LL (3.2-fold). Thus, it is also unlikely that *LhcB3* expression patterns alone can account for the observed NPQ phenotypes. None of the nuclear gene families encoding subunits of PSII exhibited expression differences between T2 and D3 exceeding 1.3-fold. Relative to D4, several subunits (*PsbW*, *PsbX* and *PsbY*) were >1.5-fold more highly expressed in T2. Again, however, combined gene family expression for *PsbW* and *PsbX* decreased from LL to EL in all three species, making it unlikely that they account for the observed increases in NPQ. *PsbY* expression

decreased significantly in D3 and D4, but not in T2, from LL to EL. It is believed that PsbY functions to stabilize Ca<sup>2+</sup> at the oxygen evolving complex, and *Synechocystis psby* mutants exhibit elevated levels of photoinhibition when grown under high light (Neufeld et al. 2004). Thus, differences in PsbY abundance under EL could possibly explain the higher levels of photodamage (qI<sub>D</sub>) in D3 and D4 compared to T2.

Recent studies have shown that NPQ is impaired in mutants with reduced levels of galactolipid synthases (MGD and DGD) (Aronsson et al. 2008; Holzl et al. 2009), and the genes encoding these enzymes emerged from our transcript profiling data as promising candidate genes to explain the observed increase in NPQ in T2. Members of both gene families responded to EL with increased expression in T2, and were up-regulated in T2 relative to D3 or D4.

Consistent with the observed up-regulation in T2 of both B-type MGD genes and DGD genes, T2 exhibits the highest level of DGDG of the three species under both conditions. In contrast, D3 exhibits the highest levels of MGDG under both LL and EL, which is in accordance with higher A-type MGD expression in this species relative to D4 or T2.

In *Arabidopsis*, A-type MGD (MGD1) is responsible for the bulk of MGDG synthesis. B-type MGDs (MGD2 and MGD3), in contrast, are principally expressed in non-green tissue in response to phosphate starvation (Benning and Ohta 2005). Under phosphate limitation, DGDG accumulates in non-plastidic membranes, and is thought to substitute for phospholipids, which can subsequently be catabolized in order to liberate phosphate. Thus, B-type MGDs are thought to be conditionally required to produce MGDG for DGDG synthesis to support remodeling of extra-plastidic membranes.

Intriguingly, in contrast to *Arabidopsis*, B-type MGDs are expressed in green leaf tissue in D3, D4 and T2, and even expressed at a higher level than A-type MGDs

in T2 under EL (Figure 4.7). This high level of B-type MGD expression in T2 correlates with a very high level of DGDG in leaves (Figure 4.9). Whereas DGDG typically constitutes ~15-20% of plastid lipid (Holzl et al. 2009), it represents 46.7% in T2 under LL. Even in D3 and D4 leaves, where B-type MGD expression is lower than in T2 (but higher than in *Arabidopsis*), DGDG represents >30% of total plastid lipid. These are similar levels to those observed in *Arabidopsis* under phosphate deficiency. It is unlikely, however, that these plants were experiencing phosphate deficiency. First, all plants in this study were fertilized regularly with NPK fertilizer. Second, under phosphate shortage DGDG replaces the phospholipid, phosphatidyl choline (PC), in non-plastidial membranes, resulting in a decrease in the proportion of PC to total lipid, but T2, which has the highest DGDG content, also has a lower DGDG/PC ratio than either D3 or D4.

Combining the expression and lipid profile data, it appears that in *Glycine* the bulk of MGDG incorporated directly into plastid membranes is synthesized by A-type MGDs, whereas MGDG used to feed DGDG biosynthesis is produced by the action of B-type MGDs. By up-regulating both B-type MGD expression and DGD expression, T2 achieves comparatively high levels of DGDG. By up-regulating A-type MGD expression, D3 achieves comparatively high levels of MGDG.

In *Arabidopsis*, B-type MGDs are believed to be associated with the chloroplast outer membrane, in close association with DGDs (which are also in the outer membrane), and specifically function in channeling MGDG into DGDG synthesis (Benning and Ohta 2005). Thus, it would make sense that high levels of B-type MGD combined with high levels of DGD in T2 would result in elevated DGDG.

These data suggest that B-type MGDs function in equivalent ways in *Glycine* and *Arabidopsis*, but are utilized in different tissues and under different conditions. *Arabidopsis* is a C16:3 plant (meaning it utilizes DAG derived from the plastid as well

as from the ER for galactolipid biosynthesis), whereas *Glycine* species are C18:3 plants (meaning they only utilize ER-derived DAG for galactolipid biosynthesis). The observed differences in B-Type MGD function may, therefore, represent a fundamental difference between C16:3 plants and C18:3 plants. It would be interesting to determine whether all C18:3 plants express B-type MGDs in green tissue, and whether all C16:3 plants restrict B-type MGD expression to specialized conditions in non-green tissues, as well as to determine which strategies are ancestral and which are derived by placing them in a phylogenetic context.

*Arabidopsis dgd* mutants exhibit reduced PSII quantum efficiency, and greater photoinhibition (qI) than wild type plants under excess light (Holzl et al. 2009). This suggests that DGDG is required for optimal functioning of NPQ processes. DGDG is found within PSII and LHCb protein complexes (Loll et al. 2005), and is thought both to stabilize protein-protein interactions and to provide a “lubricant” for protein movement within these complexes (Holzl et al. 2009). For example, four DGDG molecules surround the PSII reaction center proteins, and are thought to facilitate their removal and replacement during the PSII damage-repair cycle (Loll et al. 2005). qE and qI<sub>p</sub> are both thought to involve conformational changes in PSII and light harvesting proteins that facilitate thermal dissipation of excitation energy (Horton and Ruban, 2005; Niyogi et al., 2005). DGDG molecules may play important roles in facilitating these conformational changes. Having increased capacity for DGDG biosynthesis may enable T2 to maintain the optimal number of DGDG molecules in association with PSII and LHCb to maximize qI<sub>p</sub> under EL, when lipid turnover is high.

However, though T2 exhibits a greater proportion of DGDG than either diploid, the ratio of DGDG/total lipid decreases under EL, making it unlikely that the NPQ capacity is a simple function of DGDG abundance. Alternatively, enhanced NPQ

capacity in T2 may relate to MGDG. All three species respond to EL by increasing the fraction of MGDG in total plastid lipids (Figure 4.9). Though T2 has less MGDG per total lipid than D3, it exhibits the largest relative increase in MGDG content from LL to EL (a 23.2% increase in the fraction of total plastid lipid compared to 13.9% for D3 and 13.8% for D4). *Arabidopsis mgd1* mutants exhibit severely impaired NPQ capacity under excess light (Aronsson et al. 2008). This was attributed to increased conductivity of the thylakoid membrane. During exposure to excess light, this higher conductivity impaired the capacity to form a pH gradient across the thylakoid membrane, which is required for activation of PsbS, as well as conversion of violaxanthin to zeaxanthin. This impairment was not observed under low light (comparable proton motive forces were observed between *mgd1* and wild type at light intensities up to  $200 \mu\text{molm}^{-2}\text{s}^{-1}$ ), and it was suggested that elevated conductivity under high light results from an interplay between structural changes in PSII-LHCII complexes and MGDG content. Thus, as with DGDG, NPQ capacity does not appear to be a simple function of MGDG abundance. Rather, the roles played by galactolipids in optimizing NPQ are probably multifaceted and dynamic. By up-regulating the B-type MGD/DGD pathway, the T2 tetraploid may be better able to fulfill these roles.

## REFERENCES

- Anwaruzzaman M, Chin BL, Li XP, Lohr M, Martinez DA, Niyogi KK. 2004. Genomic analysis of mutants affecting xanthophyll biosynthesis and regulation of photosynthetic light harvesting in *Chlamydomonas reinhardtii*. *Photosynthesis Res* 82:265-276.
- Aronsson H, Schottler MA, Kelly AA, Sundqvist C, Dormann P, Karim S, Jarvis P. 2008. Monogalactosyldiacylglycerol deficiency in arabidopsis affects pigment composition in the prolamellar body and impairs thylakoid membrane energization and photoprotection in leaves. *Plant Physiol* 148:580-592.
- Audic S, Claverie J. 1997. The significance of digital gene expression profiles. *Genome Res* 7:986-995.
- Avenson TJ, Cruz JA, Kramer DM. 2004. Modulation of energy-dependent quenching of excitons in antennae of higher plants. *Proc Natl Acad Sci* 101:5530-5535.
- Awai K, Maréchal E, Block MA, Brun D, Masuda T, Shimada H, Takamiya K, Ohta H, Joyard J. 2001. Two types of MGDG synthase genes, found widely in both 16:3 and 18:3 plants, differentially mediate galactolipid syntheses in photosynthetic and nonphotosynthetic tissues in *Arabidopsis thaliana*. *Proc Natl Acad Sci* 98:10960-10965.
- Benning C, Ohta H. 2005. Three enzyme systems for galactoglycerolipid biosynthesis are coordinately regulated in plants. *J Biol Chem* 280:2397-2400.
- Demmig Adams B, Ebbert V, Mellman DL, Mueh KE, Schaffer L, Funk C, Zarter CR, Adamska I, Jansson S, III WWA. 2006. Modulation of PsbS and flexible vs sustained energy dissipation by light environment in different species. *Physiol Plantarum* 127:670-680.
- Demmig-Adams B, Adams WW. 2000. Photosynthesis: Harvesting sunlight safely. *Nature* 403:371-374.

- Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. 2004. Evolution of the perennial soybean polyploid complex (*Glycine* subgenus *Glycine*): A study of contrasts. *Biol J Linn Soc* 82:583-597.
- Holz G, Witt S, Gaude N, Melzer M, Schottler MA, Dormann P. 2009. The role of diglycosyl lipids in photosynthesis and membrane lipid homeostasis in *Arabidopsis*. *Plant Physiol* 150:1147-1159.
- Horton P, Ruban A. 2005. Molecular design of the photosystem II light-harvesting antenna: Photosynthesis and photoprotection. *J Exp Bot* 56:365-373.
- Horton P, Johnson MP, Perez-Bueno ML, Kiss AZ, Ruban AV. 2008. Photosynthetic acclimation: Does the dynamic structure and macro-organisation of photosystem II in higher plant grana membranes regulate light harvesting states? *FEBS J* 275:1069-1079.
- Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NWG, Couloux A, Dalwani A, Denny R et al. 2008. Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiol* 148:1740-1759.
- Kulheim C, Agren J, Jansson S. 2002. Rapid regulation of light harvesting and plant fitness in the field. *Science* 297:91-93.
- Lewis WH. 1980. Polyploidy in species populations. In: Lewis WH, editor. *Polyploidy: Biological Relevance*. NY: Plenum. p. 103-144.
- Li XP, Muller-Moule P, Gilmore AM, Niyogi KK. 2002. PsbS-dependent enhancement of feedback de-excitation protects photosystem II from photoinhibition. *Proc Natl Acad Sci* 99:15222-15227.
- Li XP, Bjorkman O, Shih C, Grossman AR, Rosenquist M, Jansson S, Niyogi KK. 2000. A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* 403:391-395.
- Li Z, Wakao S, Fischer BB, Niyogi KK. 2009. Sensing and responding to excess light. *Annu Rev Plant Biol* 60: 239-260.

- Loll B, Kern J, Saenger W, Zouni A, Biesiadka J. 2005. Towards complete cofactor arrangement in the 3.0Å resolution structure of photosystem II. *Nature* 438:1040-1044.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509-1517.
- Maxwell K, Johnson GN. 2000. Chlorophyll fluorescence - a practical guide. *J Exp Bot* 51:659-668.
- Melkonian J, Owens TG, Wolfe DW. 2004. Gas exchange and co-regulation of photochemical and nonphotochemical quenching in bean during chilling at ambient and elevated carbon dioxide. *Photosynthesis Res* 79:71-82.
- Miller NA, Kingsmore SF, Farmer AD, Langley RJ, Mudge J, Crow JA, Gonzalez AJ, Schilkey FD, Kim RJ, van Velkinburgh J et al. 2008. Management of high-throughput DNA sequencing projects: *Alpheus*. *J Comput Sci Syst Biol* 1:132-148.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5:621-628.
- Neufeld S, Zinchenko V, Stephan DP, Bader KP, Pistorius EK. 2004. On the functional significance of the polypeptide PsbY for photosynthetic water oxidation in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *Mol Genet Genom* 271:458-467.
- Niyogi KK. 2000. Safety valves for photosynthesis. *Curr Opin Plant Biol* 3:455-460.
- Niyogi KK, Li XP, Rosenberg V, Jung HS. 2005. Is PsbS the site of non-photochemical quenching in photosynthesis? *J Exp Bot* 56:375-382.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* 34:401-437.

- Pfaffl MW, Horgan GW, Dempfle L. 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucl Acids Res* 30:e36.
- Pfeil BE, Craven LA, Brown AHD, Murray BG, Doyle JJ. 2006. Three new species of northern Australian *Glycine* (Fabaceae, Phaseolae), *G. gracei*, *G. montis-douglas* and *G. syndetika*. *Australian Syst Bot* 19:245-258.
- Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM. 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* 339:62-66.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178-183.
- Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot* 96:336-348.
- Szabo I, Bergantino E, Giacometti GM. 2005. Light and oxygenic photosynthesis: Energy dissipation as a protection mechanism against photo-oxidation. *EMBO Rep* 6:629-634.
- Warner DA, Edwards GE. 1993. Effects of polyploidy on photosynthesis. *Photosynthesis Res* 35:135-147.
- Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873-881.