THE EVOLUTION OF GENE AND GENOME DUPLICATION IN SOYBEAN

by

BRIAN D. NADON

(Under the Direction of Scott A. Jackson)

ABSTRACT

Duplication of DNA is one of the prime drivers of diversification, speciation, and adaptation for life on earth. Plants are highly tolerant of polyploidy or whole-genome duplication (WGD) - indeed, all characterized plant genomes show traces of a history of polyploidy. Duplicated genes often diverge in their function post-duplication, and can assume subsets of their old functions, take on new functions, or be deleted altogether. Studying how these processes have affected crop plants can illuminate their evolution and show how they might be improved through breeding in the future. Legumes, and especially soybean (*Glycine max* L.), offer a valuable system to study this. For this work, first, the soybean genome was aligned to itself, which showed that gene pairs from the most recent duplication event in soybean have maintained similar expression profiles within tissues and have maintained their methylation status more consistently than older duplicate pairs. Next, an algorithm (TetrAssign) was developed to reconstruct and phase ancient soybean subgenomes post-WGD, and comparison of these reconstructions with maize showed that soybean's ancient subgenome sets were less divergent in their gene deletion, expression, and methylation profiles than maize's. Then, a set of gene families (orthogroups) for soybean and several other sequenced legume genomes was analyzed to reveal that rates of stochastic gene duplication were low, while gene deletion (death)

rates were higher but variable among the legume branches, and furthermore found that the *Glycine*-specific duplication event had a much higher retention of gene duplicates post-WGD than the *Faboideae* duplication. Lastly, resequencing of elite and wild (*Glycine soja*) soybean accessions determined that while duplicated genes dominate the gene set of soybean, orphan genes and dispensable genes are overrepresented among genes most strongly selected for during the domestication of soybean. The results of this work indicate that soybean's genome is unusually duplicated for a diploidized paleotetraploid, that its subgenomes 8-13 My ago were less diverged than was initially thought, and that the evolution of duplicate genes is ongoing in soybean and has probably impacted the transformation of soybean from a wild, vine-like plant into the dependable economic powerhouse it is today.

INDEX WORDS: soybean; legume; genome; evolution; polyploidy; epigenomics; domestication; pangenome

THE EVOLUTION OF GENE AND GENOME DUPLICATION IN SOYBEAN

by

BRIAN D. NADON

B.S., Arizona State University, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

THE EVOLUTION OF GENE AND GENOME DUPLICATION IN SOYBEAN

by

BRIAN D. NADON

Major Professor:	Scott Jackson
Committee:	Zenglu Li
	Peggy Ozias-Akins
	Robert Schmitz
	James Leebens-Mack

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

ACKNOWLEDGEMENTS

I owe many thanks to many people for my time at the University of Georgia, only a few of which I can list here. I owe my gratitude to Chunming Xu for being my serendipitous advisor on all things bioinformatics, and for just so happening to be working on the exact same hypothesis as I was, sitting directly behind me for months, unbeknownst to both of us, which led to the most fruitful collaborations I had during my studies. I owe deep thanks to Wayne Parrott for being like a second advisor to me, pushing me to be my best when I needed it the most. My thanks to my committee members for continually keeping me on the right track, and for their valuable input in my development as a scientist. Last, and most importantly, my deepest gratitude goes to all my family and friends who supported me through the most challenging four and a half years of my life.

TABLE OF CONTENTS

Page

CHAPTER 1

THE POLYPLOID ORIGINS OF CROP GENOMES AND THEIR IMPLICATIONS: A CASE STUDY IN LEGUMES[1]

(INTRODUCTION AND LITERATURE REVIEW)

Abstract: Gene duplication and polyploidy are some of the most important, yet underappreciated, evolutionary forces that have shaped all flowering plants on earth, and the crop plants that enable human economic activity are prime exemplars of duplication in action. Polyploidy involves an immediate doubling, tripling, or more of a genome, often followed by drastic chromosomal reorganization or a reduction back to diploidy and has occurred many times in the history of most characterized plant genomes. Understanding how duplication shapes plant genomes is critical for understanding plant genetics and breeding. Of particular importance are legumes, one of the largest plant families on earth, often noted for their nitrogen fixation abilities and high nutritional value due to their protein content. Among these Papilionoid (*Faboideae*) legume crops are alfalfa, soybean, peanut, and common bean. All of these have experienced polyploidy events somewhere in their history, some ancient (60 My or more) and some very recent (e.g. ~10,000 years ago in peanut). The modes by which these polyploidies arose, whether from divergent genomes coming together (allopolyploidy), or identical or similar genomes duplicating (autopolyploidy), can affect their evolution, domestication, and improvement considerably, whether by generating new functional diversity or driving speciation. Appreciating the indelible mark polyploidy and duplication leave on these legume genomes will

[1] Modified from submission to *Advances in Agronomy*.

enable a better understanding of the molecular biology, breeding, and agronomy of these critical crops.

BACKGROUND: THE SIGNIFICANCE OF POLYPLOIDY AND DUPLICATION

Polyploidy: an overview

The landscape of modern genetics and genomics, with the explosion of discoveries enabled by continuous advances in field studies, laboratory techniques, sequencing technologies, and the exponential growth of computing power, offers a staggering wealth of data and insights yet unseen in biology. While the breadth and depth of information offered by these modern techniques has allowed for new discoveries in genetics, these advances are framed, contextualized, and enabled by the fundamentals of cell biology and cytology established in large part centuries ago. Beginning perhaps as far back as the 17th century, when Robert Hooke recorded his observations of the cells of a cork tree, the importance of the smallest physically observable components of an organism was understood. With the discovery of the nucleus in the early 19th century by Robert Brown (Oliver, 1913), and the first observations of the chromosomes within that nucleus later in the same century, the foundations of molecular genetics were laid. Simultaneously, Gregor Mendel, now a household name, quietly carried out his essential work in inheritance in pea plants. The rediscovery of Mendel's work in the early 20th century, and the synthesis of cellular and nuclear biology with Mendel's laws of inheritance led T.H. Morgan and his colleagues to deduce that chromosomes, one of those essential microscopic cellular components, carried genes, and that genes were the fundamental unit of inheritance in living organisms (Morgan et al., 1922). Thus, it was clear early on that the small bundles of DNA in the nucleus, or chromosomes, and their behavior were essential to understanding biology.

Abnormalities in the expected behavior of chromosomes were noted early in the history of cytology. While it was generally understood that most sexually reproducing organisms that

were studied were of a diploid nature, having two copies of their chromosome set (one from each parent), oddities with chromosome counts and pairing were observed that suggested that this was not always the case. Of interest is polyploidy, the state of having more than 2 copies of a chromosome set in a sexually reproducing organism. Perhaps the first example of the discovery of polyploidy was in 1907, when Lutz noted that some mutants of her *Oenothera lamarckiana* plants had enlarged cells with approximately double the normal chromosome number, and were consistently inheriting this peculiarity (Lutz, 1907). Since then, polyploidy has been recognized as an important phenomenon in biology, evolution, and genetics. While it is often assumed that sexually reproducing organisms have 2 copies of their genome (diploidy), changes in chromosome numbers and sets like aneuploidy (uneven or incomplete sets of chromosomes), haploidy (a reduction in a whole chromosome set), and polyploidy are frequent across the tree of life. From polyploid yeast, to haploid male ants and bees, to tetraploid varieties of the common goldfish, to the octoploid strawberry, changes in ploidy level and chromosome number are a surprisingly common feature of life on Earth. However, polyploidy seems to be unusually frequent and tolerated in plants as compared to other multicellular eukaryotes like animals and fungi (Liu et al., 2016). It has been estimated recently via a meta-analysis that about 24% of extant plant species are polyploids (or neopolyploids, discussed later), with 76% being diploids (Barker et al., 2016). Further, an estimated 15% of angiosperm speciation events are due to ploidy changes (Wood et al., 2009). For comparison, only a single mammalian species is believed to be stably polyploid, the plains viscacha rat (*Tympanoctomys barrerae*) (Gallardo et al., 1999). Other animals like some fish and amphibians, as well as single-celled eukaryotic organisms, have also been shown to be polyploid (Mable, 2004). The frequency of polyploidy in

plants underscores that understanding the mechanisms that give rise to polyploid plants is critical for appreciating its role in evolution, and in particular, of crop plants.

While many plant scientists and geneticists acknowledge and appreciate both the frequency and importance of polyploidy, the general mode of formation of polyploid plants is often misunderstood. A model like Fig. 1.1a is often presented as the mechanism of polyploid formation (in this case, an allopolyploid), where two different species are crossed to create a diploid hybrid, whose genome then spontaneously duplicates to create a tetraploid. The process of spontaneous doubling in nature is in reality extremely rare, with scant few confirmed examples of its occurrence (e.g. *Primula kewensis*) (Newton and Pellew, 1929). In actuality, polyploid formation is a flipped version of this process (Fig. 1.1b). In this model, 2n or unreduced gametes (gametes with the sporophytic chromosome number) are produced by the parent plants, which then fertilize each other and become a polyploid individual. In many cases, the process goes through a 3x intermediate, which can then hybridize to a diploid parent in the population to form a tetraploid via a "bilateral" pathway (Fig 1.1c). The regrettably less accurate "hybridization followed by doubling" hypothesis was first popularized in 1917 with Ø. Winge's work on yeast (Winge, 1917), and while groundbreaking in its time, this conception of polyploidy remains in the minds of many biologists today. This is despite subsequent research documenting that the overwhelming mode of formation of polyploids is via 2n gamete formation (Harlan and deWet, 1975; Mable, 2004; Mason and Pires, 2015; Ramsey and Schemske, 1998).

While triploid individuals arising from a reduced-unreduced gamete cross from a 2n gamete producing population (e.g. 2x-x cross) are sterile, matched gametes (x-x, 2x-2x) would be fertile with individuals of the same cytotype and reproductively isolated from individuals of other ploidies. Thus, it is has been proposed that polyploidy is an important source of speciation

events for plants (Mason and Pires, 2015). Furthermore, there is evidence that unreduced gamete formation is not only under genetic control, but is unusually common in interspecific diploid crosses – precisely the kind of cross that produces allotetraploids. Specifically, in a species with a genome "AA" and another with "BB", crosses between the two will often show AAB, ABB, and AABB offspring much more often than would be expected (Heyn, 1977). In addition, it has been demonstrated in many cases that unreduced gamete formation is increased under stressful or extreme environmental conditions, indicating that formation of polyploids is possibly an evolutionary strategy to increase adaptive variation in a population in a short period of time (Mason et al., 2011). It is clear that polyploidy is an important evolutionary strategy for these plants and thus warrants study as a central driver of evolution in plants.

Autopolyploidy and allopolyploidy: a continuum

Historically, polyploidy has been categorized into two types, depending on the similarity of the subgenomes (the distinct, complete paired sets of chromosomes that make up a polyploid genome) that comprise the polyploid genome: autopolyploidy, which implies that the subgenomes that make up the polyploid genome are identical or come from the same parent species, and allopolyploidy, which implies that the subgenomes that make up the genome are from different parent species (Kihara and Ono, 1926; Stebbins, 1947; Stebbins, 1950). The general understanding is currently that perhaps autopolyploids are slightly more common among plants than allopolyploids, or at least are generated more frequently (Barker et al., 2016). As the names imply, it is presumed that autopolyploids arise from duplication events within an individual or within a species, such that multiple copies of a single genome exist in a single nucleus, and that allopolyploids arise from hybrids between two different species such that two

different genomes coexist in a single nucleus. The different types of polyploids thus tend to have different characteristics and genomic behavior based on their origin and mode of formation.

Many general trends and characteristics of autopolyploids and allopolyploids (or organisms closer to each respective side of the pairing continuum) have been noted over the years. Autopolyploids, having arisen from identical genomes or extremely closely related genomes, often show evidence of tetrasomic inheritance. That is, unlike their diploid counterparts, chromosomes in autopolyploids often form multivalents (e.g. quadrivalents), and thus the genetic ratios of traits are fundamentally altered as more allelic combinations are possible. For example, a trait with two alleles, 'A' and 'a', would have the combinations AA, Aa, and aa in a diploid, but in an autotetraploid, could be AAAA, AAAa, AAaa, Aaaa, and aaaa. Furthermore, in heterozygous crosses, the ratios of offspring are vastly different. In a diploid heterozygote of the same trait, a self-cross would result in 25% recovery of a recessive trait, but for an autotetraploid, AAaa selfing results in a recessive trait recovery of just 2.8%. Also, if a trait contains 3 or 4 different alleles at a locus which contribute to vigor, this vigor is always lost upon the first generation of selfing in autotetraploids (Busbice and Wilsie, 1966). Thus, autotetraploids are highly sensitive to inbreeding depression, and for this reason autotetraploid plants tend to be outcrossers (Brown, 1993; Galloway et al., 2003; Wu et al., 2001). Allopolyploids, on the other hand, arise from two divergent genomes. This means that pairing and recombination between homoeologs (chromosomes from different progenitor genomes) is presumed to be limited in most cases.

In general, the separation of homoeologous subgenomes in allopolyploids means that, in a sense, heterosis is maintained as a "built-in" feature of allopolyploid genomes. The exchange of alleles between the subgenomes is restricted, meaning that the presence of several different

allele combinations can be maintained over generations (Barcaccia et al., 2003; Herrera et al., 2002; Ising, 1966; Thomas and Waines, 1984). Thus, the genetics of allotetraploids in this regard means that they tend to be selfers (Lande and Schemske, 1985; Njiokou et al., 1993; Tsuchimatsu et al., 2012). There are exceptions to the expected inheritance patterns of allopolyploids in many cases, however, and observations of recombination and pairing between homoeologs in allopolyploids are numerous. A canonical example is the discovery of pairing genes in wheat, which allows homoeologous chromosomes to pair and recombine freely when mutated or knocked out, indicating that allopolyploids do not always strictly discriminate between subgenomes and that pairing is not necessarily bound to sequence or structural homology between chromosomes (Dvorak and Lukaszewski, 2000; Griffiths et al., 2006; Mikhailova et al., 1998; Okamoto, 1957; Sears, 1976). Although the existence of different classes of polyploid based on their pairing behavior were recognized early on, with the terms "tetraploidy" and "double diploidy" (Blakeslee et al., 1923) proposed for what might now be called "autotetraploidy" and "allotetraploidy" (Kihara and Ono, 1926), a newer, more nuanced view has arisen that challenges the strict distinction between two mutually exclusive categories of polyploids (Gaut and Doebley, 1997; Jackson, 1976, 1982; Sybenga, 1996).

Autopolyploidy and allopolyploidy, while useful categorizations, often fail to capture the breadth of observed behavior of chromosomes in plant genomes. For instance, what about polyploids (tetraploids in this example) where chromosome pairing is not simply bivalent (allo) or quadrivalent (auto) (Lloyd and Bomblies, 2016) but somewhere in between, containing bivalents, trivalents and quadrivalents? Furthermore, if these polyploid categories are defined by the divergence of the progenitor genomes that gave rise to the polyploid, a scenario can be conceived where the two progenitor genomes were sufficiently diverged to avoid autopolyploidy

but not enough to be completely allopolyploid (disomic). Thus, at the very least, a third, middle category is required, often referred to as segmental allopolyploidy (Gaut and Doebley, 1997; Jackson, 1976, 1982). In this case, chromosomes in meiosis will form both bivalents and multivalents such that loci exhibit both disomic and polysomic inheritance.

Using chromosome pairing as evidence to validate segmental allopolyploids is complicated, however, because even strict autotetraploids may not pair exclusively as quadrivalents (instead showing no preference for which homologs of a chromosome will pair), and because a strict allopolyploid with a single translocation would result in at least one multivalent (Sybenga, 1996). Although chromosome pairing may make the identification of segmental allopolyploids difficult, DNA sequencing has offered further evidence of their prevalence as in cotton (Guo et al., 2014) and peanut (discussed later), especially in cases of allele exchange between subgenomes and allele conversion (i.e. unequal allele exchange) between subgenomes (Wang et al., 2017b). Furthermore, the discovery of pairing genes and genetic control of homoeologous pairing has further blurred the line between auto and allo-polyploidy as defined by chromosome pairing patterns, since even chromosomes that are quite diverged can be made to pair in the absence of pairing control genes in wheat (Chen et al., 1994; Sears, 1976). In general, then, it is important to conceive of classes and types of polyploids not as strict classifications but rather as lying on a continuum with two extremes.

Ancient polyploidy events, or Paleopolyploidy

Although many plants and other eukaryotes are recent polyploids (‘neopolyploids’), all characterized flowering plants and many animals and other eukaryotes also show a history of ancient polyploidy events, or paleopolyploidy. Neopolyploids appear to have a tendency to reorganize and revert to a diploid mode of chromosome pairing and inheritance over long periods

of time (Ma and Gustafson, 2005; Wang et al., 2005; Wolfe, 2001), a process referred to as diploidization. The mechanisms driving diploidization are still largely a mystery, but nonetheless paleopolyploidy is ubiquitous throughout life on earth.

The critical moment for the recognition of ancient polyploidy and duplication as important contributors to evolution was perhaps Susumu Ohno's 1970 work *Evolution by Gene Duplication.* At the time of its publication the work was largely considered to be of little interest outside the field of fish cytogenetics, but later work in the emerging fields of genomics and genome comparison in the lead up to the turn of the millennium re-contextualized Ohno's work, and reaffirmed the idea that genome duplication was far more important than had been originally assumed. Early work in humans (as in the HOX gene cluster), flies, and other animal species indicated that the chromosomes of these species had long duplicated segments with homology to other segments in the genome (Adolph, 1991; Ohno, 1970; Ohno, 1973), indicative of ancient duplication events or polyploidy that had since been rearranged. Later work in maize and other grasses in comparative mapping indicated that this phenomenon was also widespread in plants, as was later further corroborated by the first whole genome DNA sequence assemblies of major plant genomes like maize and *Arabidopsis* (2000; Gaut, 2001; Schnable et al., 2009).

Because plants tolerate polyploidy, and as polyploidy can drive speciation events, evidence of paleopolyploidy in plant genomes abounds. Genome sequencing, assembly, and analysis of maize, rice, grape, brassica, potato, soybean, and many more plants have demonstrated that paleopolyploidy is a universal trait of crop plant genomes (International Rice Genome Sequencing Project, 2005; Parkin et al., 2014; Schmutz et al., 2010b; Schnable et al., 2009; The French–Italian Public Consortium for Grapevine Genome Characterization, 2007; The Potato Genome Sequencing Consortium, 2011). Furthermore, indications that duplicated genome

segments are can be characteristic of plant genomes was noted early in cytogenetic research. For example, evidence of the paleotetraploid origin of maize was described in the early 20th century via chromosome pairing behavior (McClintock, 1930; Ting, 1966) and later confirmed with isozyme, mapping, and sequencing data (Bolot et al., 2009; Devos, 2005; Gaut, 2001; Goodman et al., 1980; McMillin and Scandalios, 1980). With DNA sequence data, paleopolyploidy is typically detected via analysis of synteny, which is a state in which homologous genes are found in a similar order within or between a genome (also referred to as collinearity). Synteny analysis is a powerful tool for unraveling the evolutionary history of plants, as observing syntenic regions detected via sequence homology and collinearity can reveal sequential paleopolyploidy events. The availability of genome sequences for many plant species allows for aligning whole genomes of related species, searching for regions of homology, grouping these together into regions of synteny (syntenic blocks), and using these relationships to estimate the evolutionary histories that link these species. This method has been used to infer that all core eudicots arose from a common ancestor with a karyotype of n=7 that experienced a hexaploidy (or two coincident tetraploidies) about 130 million years ago, and that monocots arose from an n=5 ancestor which experienced a tetraploidy event 50-60 million years ago followed by two chromosome fusions (Argout et al., 2010; Salse et al., 2009).

Mechanisms and consequences of ancient genome duplications

The general patterns of genome evolution after polyploidy and diploidization are still under scrutiny, but there is no clear universal trend that defines what happens to a diploidizing genome, with different observations in different species. Some preliminary generalities have been described, however. It has been noted that tetraploidy and diploidizations are nearly always followed by extensive genome reorganization (Kasahara et al., 2007; Mandáková et al., 2010;

Parkin et al., 2014; Santos et al., 2003; Schnable et al., 2011; Wang et al., 2017a). For example, a study of *Brassica oleracea* (e.g. kale, cabbage, broccoli) showed that a hexaploidy event followed by diploidization and subsequent reshuffling of the genome gave rise to the modern *Brassica* species used extensively in agriculture today. Dozens of different chromosomal translocations or other changes since *B. oleracea*'s divergence from *Arabidopsis thaliana* resulted in 72 distinct ancestral genome blocks that moved from their original orientations (Parkin et al., 2014; Schranz et al., 2006). Maize, on the other hand, has shown fewer rearrangements following a tetraploidy event from about 10 Mya. When maize orthologous blocks were aligned to its relative lacking the tetraploidy, *Sorghum bicolor,* the reconstructed ancient maize homoeologous segments were often composed of an entire maize chromosome, albeit often with inversions or deletions. In this scenario, 5 ancient homoeologous chromosome pairs could be identified as arising from a single maize chromosomes, with another 5 being composed of two or more maize chromosomes, indicating a much lower rate of interchromosomal translocation than of inversions or other intrachromosomal changes (Schnable et al., 2011). Still, many inversions, translocations, and other rearrangements were found in the maize genome, and the same seems to be a common feature of many paleopolyploid genomes. Thus, chromosomal rearrangements may be a common feature among paleopolyploid genomes, but the reasons for this and the mechanisms that drive it remain a mystery.

One question in paleopolyploid evolution is whether diploidization precedes reorganization or reorganization causes diploidization (Mandáková et al., 2010). Some evidence in *Arabidopsis* suggests that diploidization may occur prior to large-scale genome reorganization, where chromosome pairing during meiosis in established autotetraploid cytotypes of *Arabidopsis* showed an unexpected increase in bivalents (i.e. two chromosomes forming synapses during

pairing) as compared to multivalents (i.e. 3+ chromosomes forming synapses during pairing) (Santos et al., 2003). It was noted, however, that older established tetraploid lines showed more bivalent pairing than the newer colchicine-induced autotetraploid cytotypes, suggesting that perhaps genome reorganization (in the form of deletions and translocations) played a role in changing the chromosome pairing behavior of these *Arabidopsis* plants.

The prevalence of paleopolyploidy in eukaryotes raises important questions about how it affects the evolution and divergence of plant species. Again, it is well known that genome reorganization is either coincident to or presages the diploidization of a polyploid genome. Beyond simply changing the pairing and segregation behavior of chromosomes, chromosome-scale changes can affect the inheritance of traits and individual genes. The simplest kind of change is a change in genetic dosage, where extra copies of a gene lead to a change in the balance of the fundamental biochemistry of the organism, e.g. protein levels that increase the transcription of a given gene or increase the production of a particular metabolite.

On a basic level, it has been known for about a century that the gain or loss of a chromosome arm or entire chromosome, or aneuploidy, can have a dramatic impact on phenotype. A well-known example is Down syndrome, where an extra human chromosome 21 (trisomy) causes a wide array of deleterious effects (Desai, 1997). In plants, early examples of the effects of aneuploidy were demonstrated with *Datura stramonium,* where a trisomic series of the 12 chromosomes showed that extra copies of each chromosome had different effects (Blakeslee and Belling, 1924), such as larger or smaller seed capsules or seed capsules with different shapes and textures. Trisomics were also used to map genes to chromosomes and even to specific chromosome arms (McClintock and Hill, 1931; Rhoades, 1936; Young et al., 1987),

but the specific mechanism that caused genetic differences between trisomics, which presumably had extra copies of genes compared to their euploid counterparts, was largely unknown.

More recent evidence has pointed to a gene dosage or gene balance effect as the possible reason that trisomics and other aneuploids display such striking phenotypes. In this model, an abundance of a particular enzyme due to extra gene copies being translated into their functional protein products causes increased activity (or suppression of activity if the enzyme is a suppressor) and altered stoichiometry of the biochemical processes that involve these enzymes. Furthermore, this would explain the more severe phenotypic effects of monosomic cytotypes in diploids as compared to trisomics (Birchler and Veitia, 2007).

Some of the original studies of dosage effects in plants were carried out studying the alcohol dehydrogenase-1 (*Adh*) locus of maize. However, concrete evidence of a clear effect of dosage for in this linear manner (i.e. where more copies of a gene/chromosome arm leads to more dosage) was difficult to produce as increasing the copies of chromosome arms was found not to increase dosage of ADH in a predictable manner. In contrast, however, ploidy series showed a much more predictable effect, where triploids, tetraploids, and so on produced the appropriately higher proportional dose of a gene in maize (Birchler and Newton, 1981). This suggested that, for one, the ADH producing gene was downregulated by a gene on its same chromosome arm, and additionally regulators that act upon a gene can complicate the modulation of that gene's dosage, whether on a whole-genome, chromosomal, or segmental scale.

There can be complex and highly interconnected webs of regulators for genes and their products. Dosage effects of extra copies (or fewer copies) of a gene are often constrained by these regulatory mechanisms and networks. In yeast, it has been shown that as the number of protein-protein interactions increases for a given gene product, the variation that is possible in

the expression of that gene decreases (Lemos et al., 2004). This would predict that in cases of segmental or whole-genome duplication events followed by diploidization, because all loci are now present in duplicate, gene members of large regulatory networks are resistant to deletion of any given copy, since disturbance in their dosage would affect a larger number of genes and processes. And, in fact, diploidized polyploid genomes often retain primarily transcription factors and other regulatory genes in duplicate, while other gene classes return to a diploid state (Babu et al., 2004; Blanc and Wolfe, 2004; Maere et al., 2005; Seoighe and Gehring, 2004; Teichmann and Babu, 2004; Thomas et al., 2006). While drastic changes to a network (e.g. gene deletion in a crucial rate-limiting enzyme) can sometimes be catastrophic, more subtle changes in gene function or activity level can contribute to functional diversity. In essence, this means that gene regulatory networks and gene dosage effects are critical not only in understanding how genomes evolve after duplication, but in how changes to these networks and interactions after gene duplication contribute to increasing biological diversity.

Gene and chromosome level effects of ancient polyploidies are just a few of the many different scales at which we can appreciate how duplication can drive evolution. Expanding the scope to the genome or subgenome level reveals even more broad-scale and significant consequences of genome duplication. Autopolyploidy and allopolyploidy, as discussed earlier, differ both in their origins and their effects on the genetics of the organism. It stands to reason, then, that the type of polyploidy event that gave rise to a modern paleopolyploid plant would have distinct effects on the genome.

Since allopolyploid events arose from divergent progenitor genomes, it is possible in some cases to reconstruct the ancestral state of the genomes that gave rise to the modern paleopolyploid plant. In maize, for example, ancestral pre-tetraploidy chromosome states were

reconstructed via alignment to *Sorghum bicolor* chromosomes, a related grass species whose genome lacks the maize tetraploidy event of ~10 Mya. This reconstruction demonstrated that not only was it possible to reconstruct the pre-tetraploid maize subgenomes but that there were quantifiable differences between the genomes in the number of genes that were deleted or retained after diploidization in the maize genome (Schnable et al., 2011). Specifically, one subgenome had consistently more deletions of genes and lower expression levels of genes than the other. Similar patterns have been described in other species (Buggs et al., 2014), for example *Brassica oleracea,* where expression and methylation data revealed that the hexaploidy event that gave rise to modern *Brassicas* left two subgenomes that were more prone to gene deletion, had lower gene expression levels, and showed more gene body methylation (Parkin et al., 2014).

The general term for the observation that one ancient subgenome appears to retain more genes, be less methylated, and more strongly expressed than its counterpart subgenome is "genome dominance", and has been hypothesized to be a feature of paleopolyploid genomes whose progenitor duplication event was an allopolyploidy (Buggs et al., 2014; Garsmeur et al., 2014). As a corollary, the process by which genome(s) lose or modify their duplicated genes over time is termed "fractionation" (Tiley et al., 2016).

Of course, not all ancient duplication events were allopolyploidies; accordingly, there is evidence that a different type of post-polyploidy mechanism is at play in genomes derived from autopolyploidies. It should be noted that discussing the genome in terms of 'subgenomes' would be misleading, as there are no subgenomes in an autopolyploid, but instead a set of chromosomes that can form multivalents at meiosis. However, because the ploidy types exist on a spectrum as noted earlier, and because no other suitable terminology exists, we will continue to use 'subgenomes' to discuss the effects of ancient autopolyploidy on modern genomes (

"homoeologous groups" does not work either as there are no homoeologs in a true autopolyploid). Exemplified in genomes like diploid banana (D'Hont et al., 2012) and poplar (Freeling et al., 2012; Garsmeur et al., 2014), in cases where there is no discernible genome dominance nor distinguishing characteristics between the ancient subgenomes, genes are retained in duplicate at nearly equal rates across the genome, and any gene deletions, expression differences, or methylation patters tend to be stochastic on a gene-by-gene basis (Garsmeur et al., 2014).

THE IMPORTANCE OF LEGUMES

Taxonomy and evolution

Fabaceae, *Leguminosae,* or legumes, are the third largest family of flowering plants (angiosperms), comprising a wide variety of economically and scientifically important genera and species. There are almost 770 genera and more than 19,500 species in the family, and a staggering amount of diversity therein. Legume species span the globe, with representatives in nearly every biome from deserts to tropical forests. They grow as shrubs, trees, and even aquatic plants, display diverse flowering morphology, and are adapted to a wide variety of ecological and climate conditions. Legumes, therefore, represent a prime example of diversification and adaptation in plants. They are thought to have arisen about 60 million years ago, just after the K-Pg or K-T extinction event which killed the dinosaurs, when many different angiosperm lineages were expanding and flourishing (Brea et al., 2008; Bruneau et al., 2008; Lavin et al., 2005; The Legume Phylogeny Working Group, 2017).

The largest subfamily of the legumes, and perhaps the most important economically, are the *Faboideae*, previously known as the Papilionoids. This clade includes essentially all of the legumes that are popularly known or economically important, like soybean (*Glycine max*), peanut (*Arachis hypogaea*), common bean (*Phaseolus vulgaris*), alfalfa (*Medicago sativa*), pea (*Pisum sativum*), licorice (*Glycyrrhiza glabra*), cowpea (*Vigna unguiculata*), chickpea (*Cicer arietinum*), and many more. Even the pea used by Gregor Mendel to lay the foundation of genetics and most of modern biology was a legume, further underscoring their value as economic powerhouses and as tools for scientific discovery.

One particular trait that garners considerable attention for the family is nitrogen fixation. Legumes are often synonymous with the ability to fix nitrogen from atmospheric N_2 into

biologically available NH_3 in the soil via a special symbiotic relationship with rhizobia, where soil-borne bacteria establish themselves within the roots of legume plants. Nitrogen fixing legumes can yield hundreds of pounds per acre of nitrogen per year in a field crop setting. This nitrogen fixing behavior is highly valuable, as it obviates or mitigates the need for nitrogen fertilizer when growing legumes and enriches nitrogen in the soil in crop rotation schemes. Furthermore, the nitrogen fixed from the atmosphere by these plants allow them to produce high amounts of protein, since amino acid synthesis requires biologically available nitrogen. This means that legumes are extremely valuable as a source of nutritional protein for animal and human consumption. It is important to note, however, that this nitrogen fixation behavior is not monophyletic or universal among legumes. The ability to form symbiotic relationships with soil-borne nitrogen-fixing bacteria is neither unique to legumes in the *Fabaceae* or their parent nitrogen-fixing clade, nor is it universal among these taxa, as this trait has been lost or gained in many different lineages in many independent evolutionary events (Li et al., 2015). Nitrogen fixation is estimated to occur in about 88% of legume species, underscoring it as a typical yet far from omnipresent trait for the clade (Faria et al., 1989).

Economics and agronomy

Approximately 12 to 15% of the earth's arable land area (180 million Ha) is used to grow legume crops, which account for about 27% of world crop production (Graham and Vance, 2003; Vance et al., 2000). Legumes also produce 35% of the world's processed vegetable oil. With a current world population of about 7 billion people, and rapid growth (albeit slower than historical rates) projected to reach approximately 9 or 10 billion people by 2050 (DeSA, 2013; Lutz et al., 1997), meeting the nutritional needs of a rapidly growing human population will be a significant challenge for policymakers, farmers, plant breeders, agronomists, and scientists from

all disciplines to face. Legume crops will play a crucial role in facing these challenges in the coming years and beyond, both as food for direct human consumption and as animal feed.

Legumes are excellent direct nutritional sources of protein, with the nitrogen afforded to them via nodulation allowing them to produce high amounts of protein that is stored in the seeds of the plant (Mosse, 1990). Plants still dominate as the primary source of protein for the world, with about 65% of global protein calories consumed coming from plant sources (Grigg, 1995). Legumes, then, can offer a significant source of dietary protein for the world, so much so that they are often pejoratively referred to as "the poor man's meat". Common bean and cowpea, for example, can often have 25% or more of their seed weight comprised of protein (Baptista et al., 2017), demonstrating that legumes are in many cases nearly on par with animal meat in terms of protein by weight. Furthermore, legumes are less expensive to produce and bring to market than animal products, with legumes and dry beans costing about \$0.18 $kcal^{-1}$ and meat or poultry costing about \$0.41 $kcal^{-1}$. Put simply, legumes are an efficient, valuable, and stable source of nutrition that will be essential to feed a growing population.

Earth's human population is growing quickly, but importantly so is its demand for animal protein, especially among the most populous nations (Delgado, 2003). In fact, many projections indicate that global demand for animal food products will double by 2050, although consumption of meat is much higher in developed nations than in lesser-developed nations. Raising livestock to meet this demand requires feed, and especially feed that is high in protein, like soybean meal. Thus, legumes are essential in not only meeting the plant nutritional needs of the human populace, but also in producing dairy and meat. In 2017 the United States, 12.7% of the diet of top livestock raised for food was fed with soybean meal (second only to corn at 50.3% of the

livestock diet) (AFIA 2017). Alfalfa, another legume, is grown for use as hay for livestock feed, and in 2017 the United States produced over $16B worth of alfalfa hay (USDA-NASS 2017).

Polyploidy and Genome Duplication in Legumes: A Case Study

Understanding their genetics and the structure, evolution, and behavior of legume genomes will be important in unlocking the full potential of these plants to feed the world. Despite their similarities, categorizing and separating the common crop legumes by their genomic structure and evolutionary history, especially through the lens of polyploidy (including paleopolyploidy and duplication) is useful for understanding the complexity and diversity found in the legume clade. Within the legumes, there are examples of every kind of polyploidy or duplication class: neopolyploids and paleopolyploids, and allopolyploidy and autopolyploidy. Examples of each and how it affects their genomes, genetics and breeding are discussed below.

Neopolyploid legumes

As stated before, many extant flowering plants are neopolyploids, exhibiting non-diploid inheritance, and legumes are no exception. Examples of both general classes of neopolyploids, allopolyploids and autopolyploids, can be found in crop legumes. The sheer size of the *Fabaceae* means that these examples abound, but for the purposes of this review, a few of the most economically important crop plants in the *Fabaceae* will be the primary focus. To that end, two prominent examples of each kind of neopolyploid are peanut, a neoallopolyploid, and alfalfa, a neoautopolyploid.

Peanut: a recent allopolyploid

Peanut, *Arachis hypogaea*, is a selfing, annual legume plant endemic to South America near modern-day Peru, Bolivia, or Argentina. It is an important source of oil and protein, and the most popular grain legume in Africa (FAOSTAT 2017). Worldwide, about 40 million metric

tons of peanuts were produced on 25 million hectares, and in the US alone in 2017 peanuts were a $1.1 billion market (USDA-FAS 2018). Perhaps peanut's most prominent feature is geocarpy, meaning that it produces its fruits below ground, a rare trait in plants and especially rare among crop plants. The flowers of a peanut plant open above ground, but after fertilization, the ovary, deep within the stem from which the flower arises, elongates and implants itself into the ground in a structure called a 'peg'. Eventually, this leads to the formation of podded fruit a few centimeters underground, attached to the peg (Smith, 1950). Due to peanut's geocarpy, it is commonly called "groundnut".

Modern tetraploid peanut, *A. hypogaea*, is believed to have arisen from a single allopolyploidy event between *A. ipaensis* (B-genome donor) and *A. duranensis* (A-genome donor) (Moretzsohn et al., 2013; Seijo et al., 2007), giving rise to a 2.7 Gbp allotetraploid genome denoted as "AABB" (two A genome copies and two B genome copies). Allopolyploid genomes have been postulated to arrive from hybridization of two species followed by chromosome doubling, but as stated earlier, this is unlikely and instead it is more likely that allotetraploid peanut was formed by the hybridization of two unreduced gametes from each donor species (Brownfield and Köhler, 2011). The two progenitor diploid species' genomes are estimated to have diverged around 2-3 million years ago (Moretzsohn et al., 2013), but molecular and archeological evidence suggests the tetraploid *A. hypogaea* is quite young in evolutionary terms at about 10,000 years (Bertioli et al., 2016; Moretzsohn et al., 2013).

The result of this history is that peanut's diploid progenitors are closely related to the subgenomes of tetraploid peanut. In fact, the similarity of the diploid progenitors to the polyploid subgenomes in peanut is something of an outlier among crop plants, perhaps owing to the peculiar reproductive behavior and reproductive isolation of tetraploid peanut and its

progenitors (Bertioli et al., 2016). Estimates of divergence times between the A and B tetraploid subgenomes and their respective progenitors using nucleotide substitution rates are 247,000 years and 9,400 years, indicating that the B subgenome is more closely related to *A. ipaensis* than the A subgenome is to *A. duranensis* (Bertioli et al., 2016). It is unlikely that the hybridization of the two species to give rise to *A. hypogaea* took approximately 238,000 years; instead it is more likely that the A subgenome underwent more extensive changes post-hybridization than did the B subgenome.

One genome in an allopolyploid plant undergoing significantly more extensive changes than another has been previously observed. Interestingly, this phenomenon is not confined to plants, and has been observed even in species like frog (Furman et al., 2018). The genome that is more conserved in this scenario is often called a "pivotal genome," and this terminology has been applied to other allopolyploid crops like wheat (Mirzaghaderi and Mason, 2017). In this model, the more changed genome is called the "differential genome", and the differences in how the two (or more) genomes in the polyploid genome behave after polyploidy are often explained as being a result of pre-existing differences between the progenitor diploid genomes (Kimber and Yen, 1988; Mirzaghaderi and Mason, 2017). In the case of peanut, then, the B genome (from *A. ipaensis*) would be the pivotal and the A genome (from *A. duranensis*) would be the differential (Kochert et al., 1996). This is borne out in the observation that the A genome has undergone all of the major rearrangements of the tetraploid peanut genome, and has smaller chromosomes and less DNA overall, indicating it has lost more of its genetic content. It could be said, then, that many of the common patterns in changes and evolutionary trajectories of allopolyploids and paleopolyploids could be related to pre-existing differences in the parental genomes. In peanut, the progenitors *A. duranensis* and *A. ipaensis* have many one-to-one orthologs between the two

species, but might have had significant underlying differences in their orthologous genes' expression or sequences, leading to observable differences in tetraploid *A. hypogaea*'s genome today (Song et al., 2017; Song et al., 2018). It is still unclear, however, how this older observation of "pivotal" genomes in polyploids, often based on non-sequence evidence like gel banding, correlates with genome dominance, where certain subgenomes in paleopolyploids maintain more genes than their sister subgenomes.

In addition to a marked bias in DNA loss and rearrangements that has seen the A subgenome change more in relation to its progenitor (*A. duranensis*) than the B genome (whose progenitor is *A. ipaensis*), peanut has several other interesting characteristics. One example is the observation that recombination has occurred, and continues to occur, between the A and B subgenomes in tetraploid peanut. It is expected that in a "true" allopolyploid, the subgenomes pair, recombine, and segregate independently in meiosis (i.e. A chromosomes never pair with B), but in peanut there is evidence of extensive breaking of this rule (Leal-Bertioli et al., 2015; Nguepjop et al., 2016). Marker data from mapping studies and later sequence data demonstrated that while, for the most part, peanut's subgenomes are differentiated enough to be inherited separately, there are up to about 3% of markers that have been exchanged between subgenomes – a clear sign of tetrasomic pairing and recombination (Leal-Bertioli et al., 2015). This observation has more than simply cytological implications; it means that if a locus of interest to a breeder is located on a segment of the genome that is subject to tetrasomic recombination, the segregation ratios of the trait in question could be skewed, meaning that a recessive trait that is normally not recoverable can become recoverable, or a trait that is normally recoverable in offspring can become unrecoverable. On top of potentially complicating or easing breeding efforts, non-preferential pairing of the tetraploid peanut genome indicates further that segmental

allopolyploidy is quite common and that allopolyploidy is not always what it seems. In essence, peanut is yet another exemplar that demonstrates that binary classification of types of polyploids into "allo-" or "auto-" may be misguided. Furthermore, breeding to improve peanut is complicated not only by its tetraploidy, but by the frequent interchange of loci between the subgenomes.

Alfalfa: an autopolyploid legume

In addition to allopolyploid crop plants there are autopolyploids, like potato and alfalfa. Alfalfa (*Medicago sativa*) or lucerne is an herbaceous perennial plant native to the Middle East and central Asia. It is typically grown as a grazing, forage, silage, or cover crop. It is relatively hardly and broadly adapted, grown all over the world in many different environments. As a source of feed for livestock, alfalfa is important economically; the global alfalfa hay market value was approximately $800 million in 2016 (Mordor Intelligence 2017). Alfalfa is an outcrossing species, with varying ploidies in its cultivated forms, though cultivated alfalfa is primarily an autotetraploid. Autotetraploidy has particular benefits and challenges for breeders and growers, as it can improve vigor but also complicate trait segregation ratios and accelerate inbreeding depression.

The 'gigas' effect of polyploidy, where increased ploidy increases cell and tissue size, has been well documented across plant species. Having more DNA is directly and positively correlated with cell size; specifically, doubling DNA content doubles cell volume (Abel and Becker, 2007; Müntzing, 1936; Tsukaya, 2013). This means that, when observed under a microscope, the apparent increase in cell diameter is not 2x, but rather 1.26x, since the volume has double but the radius of a cell increases by the spherical volume formula. Autopolyploidy, then, offers an opportunity for breeders of vegetative crops, since increasing ploidy directly

increases the overall size and biomass of vegetative tissue in a plant. Alfalfa, as a vegetative crop, benefits from its autotetraploidy for this reason.

The increased cell size and vigor arising from autopolyploidy does come with various drawbacks, however. For one, it has been observed most autopolyploids suffer from significant inbreeding depression after a very few generations (Parisod et al., 2010; Rausch et al., 2005). Initially it would seem that autotetraploidy would help prevent inbreeding depression by making it more difficult to accumulate collections of fully homozygous recessive traits, as four recessive alleles are required to be inherited together for this to happen, however, this is complicated by the mathematics of allele inheritance in an autotetraploid and other features of autotetraploidy like double reduction at meiosis.

Firstly, if the inbreeding coefficient *F* is defined as the probability that two alleles inherited from a progenitor are identical by descent, then doubling a heterozygous plant rapidly increases the inbreeding coefficient. If a plant has a heterozygous genotype at a locus A_1A_2, and is doubled to form an autotetraploid genotype of $A_1A_1A_2A_2$, then the probability of receiving two of the same allele from the same parent (A_1A_1 or A_2A_2 gametes) is 1/3. An inbreeding coefficient *F* of 1/3 is equivalent to about 5 or 6 generations of inbreeding in a diploid heterozygote, meaning that the simple act of becoming a tetraploid very rapidly exposes the genome to inbreeding depression, and is an issue for improvement of tetraploid alfalfa.

Meiosis and chromosome pairing also work against autopolyploids: the process of double reduction means that genetic diversity is lost at a greater rate from parent to offspring in autopolyploids than in diploids. Double reduction is the ability to recover recessive nulliplex (i.e. “aaaa”) genotypes from a triplex cross (e.g. AAAa x aaaa). This is possible because in quadrivalent pairing, crossovers between the centromere and a gene near the telomere result in

the possibility of the recovery of sister chromatid genotypes in a single gamete, meaning that two rounds of reduction in allelic diversity in the gametes are possible (normally, there is one) (Bourke et al., 2015; Wu et al., 2001).

In addition to inbreeding depression through double reduction, fertility is also a concern for autotetraploid crops (Bingham and Gillies, 1971). In an idealized autotetraploid, where all of its chromosomes pair as quadrivalents (four chromatids recombining together) in meiosis, there is a distinct reduction in fertility due to basic properties of probability. Since the separation of chromosomes after pairing in meiosis is random, a set of 4 homologous chromosomes can separate in several different ways. They can split 1-3, 2-2, or 3-1, as opposed to the simplistic, even 1-1 separation of diploids, which means that some resulting gametes will have aneuploid (i.e. incomplete or trisomic) chromosome sets. Gametes with incomplete or aneuploid chromosome sets have reduced viability, leading to a reduction in fertility for an autotetraploid (or hexaploid etc.) whose chromosomes pair in multivalents.

However, there is an important caveat: autotetraploids whose chromosomes *do not* pair in multivalents and instead pair preferentially in bivalents can avoid reduced fertility. Bivalent pairing (i.e. two chromosomes pairing together in meiosis) means that chromosomes can only split 1-1, meaning each resulting gamete will receive the appropriate chromosome set. Thus, autotetraploids with high levels of bivalent pairing are more fertile than those with multivalent pairing. Alfalfa's chromosomes pair primarily in bivalents (Armstrong, 1954), which greatly improves its fertility and ease of breeding. Although pairing occurs primarily in bivalents and not trivalents or quadrivalents in alfalfa, homologous chromosomes appeared to associate freely, indicating the true autopolyploid nature of the tetraploid species (McLennan et al., 1966). Incidentally, this demonstrates that defining the polyploidy type of a species by chromosome

pairing behavior in terms of formation of multivalents is likely inaccurate, since autopolyploids like alfalfa pair in bivalents which show no preference for which homologs they pair with (since, presumably, they are not diverged significantly enough to segregate as allopolyploids).

Paleopolyploidy in Legumes

Legumes have at least 2 shared ancient polyploidy events

Like all flowering plants, legumes have a well-documented history of paleopolyploidy (Fig. 1.2). Among Papilionoid legumes, there are at least two major detectable polyploidy events that may have contributed to diversification of the clade. All core eudicots (the clade subtending the rosids and asterids) share a putative ancient hexaploidy as revealed by studies comparing *Arabidopsis,* grape, papaya, and poplar genomes (Tang et al., 2014), referred to as the γ event. This event is presumed to be about 100-130 million years old, making it the oldest detectable whole genome duplication or paleopolyploidy event among plants (Bowers et al., 2003; Zheng et al., 2013). The ancestral eurosid genome, which experienced this hexaploidy event, is proposed to have had 7 chromosomes, giving 21 total chromosome "groups" that contributed to the evolution of modern dicot genomes (3 x 7 = 21). These ancient duplicated chromosomes are visible as degraded blocks of synteny in modern legume genomes, along with other extant eurosids.

In addition to the approximately 130 million year old eudicot hexaploidy, the Papilionoid (*Faboideae*) legume clade (which includes species like soybean, alfalfa, peanut, and common bean) has another shared ancient polyploidy event (Shoemaker et al., 2006). Early work linkage and RFLP (restriction fragment length polymorphism) maps between legumes suggested that there was a duplication event shared by legumes in the *Faboideae* clade (Shoemaker et al., 1996). Later evidence offered by BAC fingerprinting and hybridization further confirmed the

presence of an ancient duplication shared by *Faboideae*, but the age of this duplication was not discernible until sequence-based data allowed for the calculation of nucleotide substitution rates and molecular clock calibration to place the event at about 55 Mya (Shoemaker et al., 2006). Notably, this is roughly coincident with the K-T extinction event (as are many other major plant genome duplication events), lending credence to the hypothesis that duplicated genetic content can offer a potential evolutionary bulwark against succumbing to mass extinction events (Crow and Wagner, 2005; Fawcett et al., 2009; Vanneste et al., 2014a). This date does not predate the appearance of nodulation behavior among legumes, however, indicating that at the very least, nodulating behavior did not arise from duplicated copies of precursor pathway genes (Cannon et al., 2010). On the other hand, it is possible that the Papilionoid duplication did enhance the efficiency of nodulation, or allowed for more diversity in the particulars of nodulation processes (Li et al., 2013). Regardless, the ~55 Mya polyploidy event and the ~130 Mya hexaploidy likely had had significant impacts upon the evolution and diversification of legumes, and perhaps helped to ensure the survival of the ancient ancestors of modern legumes in the face of mass extinction events like the Chicxulub impact.

Specific lineages within the *Faboideae* (Papilionoid) legumes also show more recent polyploidy events, with some, like peanut, extremely recent at approximately 10,000 years old. Other events are slightly older and have since been diploidized, like the paleotetraploidy found in the genus *Glycine*, containing soybean *Glycine max,* from about 8-13 Mya (Schmutz et al., 2010b). This event is presumed to have occurred at the base of the genus, and perhaps led to the divergence of *Glycine* in the first place, since no *Glycine* species has been found that lacks this paleotetraploidy event (Pfeil et al., 2005; Ratnaparkhe et al., 2011) – though even more recent polyploidy events in *Glycine* species are found in Australia such as neopolyploid races of

Glycine tabacina (Doyle et al., 2004; Singh et al., 2001). Thus, the *Glycine* genus shows extensive duplication in their genomes because they have at least 3 major duplication events in their history, and some a fourth even more recent event. For soybean, this has very important implications in that that many genes and pathways have genetic redundancy complicating mutation breeding and functional genetics.

The effects of paleopolyploidy on modern legume genomes

While it is apparent that a whole-genome duplication or polyploidy event is perhaps the most significant, saltatory event that could happen to a genome, it is less clear how exactly these kinds of events might affect the evolution of specific traits and variation within and between species. Duplication is a source of standing variation for mutation, selection, and genetic drift to act upon, meaning that duplication is essentially an evolutionary shortcut to adding to variation. The means by which this works is explained by what is often called the "X-functionalization" framework - "X" standing for "neo-", "sub-", or "non-", all different ways to describe how duplicate gene functions can evolve (Panchy et al., 2016).

If a single gene were essential for survival, it would be evolutionarily constrained; for example, any new variants that are not immediately and directly beneficial to the fitness of the organism would be quickly selected against. Thus, many single copy genes in a genome are under purifying selection. With the presence of duplicate copies of these genes, selection on any one copy may be relaxed, as deleterious mutations would be masked or compensated by the presence of another functional copy of the gene. This allows for new mutations that can alter the function or the expression patterns of genes, or sometimes even create entirely new functions (Li et al., 2005; Lynch and Conery, 2000; Moore and Purugganan, 2005).

These new gene functions can extend, enhance, augment, or even disrupt the duplicate sister genes' functions. For example, in soybean, it has been shown that the E4 locus, a flowering time gene associated with different maturity groups, has at least two different copies owing to the recent paleotetraploidy in the *Glycine* lineage. A specific set of nonsense deletions in one copy of the E4 gene in soybean led to adaptations that allowed soybean accessions with these alleles to thrive in high latitude, long-day environments, demonstrating that the plant maintains a functional phenotype but has developed a new behavior by modifying one of its copied genes' functions (Tsubokura et al., 2013). Furthermore, it has been shown that in the very closely related wild relative of soybean, *Glycine soja,* mutations in many of the different copies of the gene family containing E4 led to a diverse range of adaptation to different climates and season lengths via modulation of maturity and flowering time (Li et al., 2014). For highly selfing species such as the *Glycine,* having multiple copies of these flowering time and maturity genes likely allowed for increased adaptive variation through mutation of multiple gene copies where admixture and recombination were scarce or absent. Similarly, divergence between paralogs of the Dt1 gene in soybean, arising from the Papilionoid polyploidy event, is suspected to be responsible for the emergence of the determinate growth habit in many different genotypes of soybean. Although both copies of the gene were highly similar in sequence, their expression profiles were different such that one copy of the gene was localized mostly to shoot apical meristems and was thus responsible for the determinate growth phenotype, while the other copy had a different, possibly unrelated function (Liu et al., 2010).

Gene duplication has also contributed to critical agronomic traits in other plants, like tomato, where a retrotransposon-mediated gene duplication and a subsequent increase in expression of one resulting gene copy (the SUN locus) led to the elongated fruit phenotype seen

in many commercial tomato cultivars (Xiao et al., 2008). In summary, it is clear that duplicate copies of genes have critical roles in plant evolution and diversification as a source of new DNA for selection, drift, and mutation to act upon to produce new genotypes and phenotypes in legumes and in plants in general.

Some general patterns and trends have been observed in plants that describe how duplicate genes tend to evolve, and the evolutionary constraints they might be under. In general, it has been observed that when genes are duplicated, their original function and relationships to other genes help determine their eventual fate. Transcription factor genes, for example, tend to be retained in duplicate within genomes at a higher rate than other types of genes. In *Arabidopsis*, it was observed that families of transcription factor genes were larger than similar gene families in other eukaryotes, and that this rate of transcription factor gene family expansion was even higher than would be expected given the high rate of genome duplication events in plant genomes (Paterson et al., 2006; Shiu et al., 2005). In soybean, it was noted that about 12.2% of protein-coding loci were transcription factors, compared to 7.1% in *Arabidopsis,* indicating that perhaps the relatively recent paleopolyploidy in soybean has led to an expansion of these apparently duplication-tolerant genes (Schmutz et al., 2010b). A more recent study has shown that genes involved in whole genome or segmental duplications in soybean and common bean tend to be transcription factors or other DNA-binding genes, while genes duplicated in other manners (e.g. tandem duplication) were enriched in other categories like ADP binding or defense pathways (Xu et al., 2018).

Taken together, it is clear that some characteristic or behavior of transcription factors allows for them to be retained in duplicate more than other types of genes. This could be explained by a corollary observation among paleopolyploid genomes that there is a tendency for

dosage-sensitive genes to more often be present in single rather than multiple copies, or to be maintained in balanced sets of copy numbers. In other words, genes that are dosage sensitive tend to maintain ancestral dosage ratios. This was demonstrated in early studies in yeast and humans, which have shown that genes that lead to fitness defects when overexpressed also tend to be genes involved in multi-protein complexes. Furthermore, it has been noted in yeast that genes in large families (>3 members) rarely encode members of multi-protein complexes, and in plants genes whose products are members of complexes tend to be resistant to pseudogenization, deletion, or nonfunctionalization (Veitia et al., 2008).

In essence, if a protein complex requires gene products "A" and "B" in an "ABA" arrangement, any deviation from gene dosage in a 2:1 ratio would be expected to cause deleterious phenotypes or an imbalance in gene products. Thus, genes involved in these types of complexes are generally retained as single copy or all in duplicate, rarely duplicating or deleting individual genes in the complex, as this would disturb the ratio of the complex. Again, the observation that in soybean and common bean, genes associated with paleopolyploid duplication events are more often transcription factors or regulatory genes (which by nature associate with DNA and other transcription factors in dosage-sensitive complexes) lends credence to this hypothesis (Xu et al., 2018).

As discussed previously, another trend in polyploid genome evolution is the observation of genome 'dominance', where one ancient homoeologous set of genes is favored in terms of expression and resistance to duplicate gene deletion. While much of the work establishing this trend was performed in studies on brassica and maize, investigations into discerning whether this phenomenon holds for ancient polyploidy events in legumes is scarce. In maize, for each pre-

maize chromosome, one particular ancient subgenome segment retains more genes and is expressed at higher levels than the other (Schnable et al., 2011). Similarly, in *Brassica oleracea*, an ancient hexaploidy left the modern *B. oleracea* with three identifiable ancient subgenome groups, where one copy is considered 'dominant' over the other two (Parkin et al., 2014).

Evidence for genome dominance in legumes, and particularly in soybean, has until recently been inconclusive or found to contrast with observations in other species. In soybean, it has been difficult to assign ancestral subgenome blocks from the 8-13 my old *Glycine* tetraploidy to specific subgenomes, mostly because of the lack of diploid ancestors (Garsmeur et al., 2014; Wang et al., 2017a). For maize and brassica, simple alignment of syntenic groups via dotplots to related species (e.g. maize segments that align to sorghum chromosomes) suffices to assign subgenome groups and reconstruct putative ancestral chromosome states pre-polyploidy. In soybean, however, this largely fails, as there have apparently been myriad chromosome rearrangements between soybean and its close relatives, e.g. common bean. Despite these limitations, in situations where these reconstructions have been attempted in soybean, there appears to be little or no bias in gene deletion levels, expression, or methylation between ancient subgenome groups in soybean (Wang et al., 2017a; Zhao et al., 2017), suggesting that perhaps soybean's ancient duplication event was under different constraints than those found in brassicas or maize, or perhaps that it was not the result of a strict allopolyploidy in the first place.

The lack of genome dominance following soybean's ancient duplication has led to the hypothesis that this event was likely an autopolyploidy or segmental allopolyploidy (Zhao et al., 2017), contradicting earlier marker and cytology-based evidence that soybean's paleotetraploidy was an allopolyploidy event (Gill et al., 2009; Schmutz et al., 2010b). Were this true, it would be expected that there is little or no bias in gene retention or expression between the ancient

soybean subgenomes, and that any gene loss would be purely stochastic and subject to the constraints of gene dosage balance. Indeed, recent studies have suggested this is the case, and that soybean's genes are not subject to large scale changes in expression or gene deletion in large syntenic blocks, but rather have diverged stochastically as paralogous gene pairs (Xu et al., 2018; Zhao et al., 2017). In essence, because soybean's genes are generally retained in duplicate, and these duplicate copies generally do not show highly divergent expression, methylation, or mutation rate patterns, most soybean genes exist in gene dosage balance with their paralogs and interacting genes. This would mean that deletion of a single copy is perhaps selected against and that any gene losses, subfunctionalization, or nonfunctionalization are purely stochastic and not the result of pre-existing differences in the ancestral genomes that gave rise to the modern paleotetraploid soybean. Again, this is in contrast to other previously studied neo- and paleopolyploid species, where often the subgenomes experience biased selection, mutation rates, expression divergence, and methylation patterning. The evidence for the nature of the duplication event that gave rise to modern soybean is still unclear, but a consensus is emerging that, at the very least, the old assumption that soybean arose from an allopolyploidy is worth re-examining in light of these observations.

The origins of the older duplication events common among Papilionoid legumes ~58 My and ~110 My old are less clear, perhaps as a result of their age. In soybean and common bean, as expected, gene paralogs arising from these events are more diverged than those arising from the newer *Glycine*-specific event 8-13 Mya, In general, the expression of a gene pair from one of these two older events is quite divergent, where within a given tissue, expression of both members of a pair from either of these legume ancient polyploidy events is only loosely correlated ($R < 0.5$) (Xu et al., 2018), indicating that these genes are diverged in function, at least

by spatiotemporal expression patterns. It is much more difficult to ascertain the nature of these duplication events compared to the *Glycine* event discussed previously, since not only have the gene copies diverged and been deleted to a much greater degree, sufficient time has passed that the differences between an auto- or allopolyploidy are greatly diminished. Also, there is no appropriate null comparison to be made to a sufficiently closely related non-*Faboid* legume species that lacks the duplication from 58 My ago or the core eudicot event 110 Mya. Regardless, future sequencing and annotation of non-*Faboideae* legume species could prove valuable for resolving what the nature of the *Faboideae* duplication event and how it shaped these genomes, especially in light of some approaches attempting to define fractionation bias in similarly old duplication events in e.g. *Brassicaceae* (Emery et al., 2018).

FUTURE PERSPECTIVES

Polyploidy, gene duplication, and the 'abominable mystery' of plants

Charles Darwin referred to the incredible diversity and range of flowering plants on Earth, their rapid expansion, and their abundance in nearly every corner of the globe as his "abominable mystery". Indeed, the stunning richness of variety and universal distribution of the angiosperms is appreciable to any who study plants. The range of adaptation of plants is so broad that many climate classification systems use the particular mix of plant species endemic to a geographic area as a defining feature of the biomes of the globe (Whittaker, 1962). Debate has stirred for centuries over how, where, and when the earliest angiosperms arose and subsequently rapidly colonized the globe. Many have pointed to the complex and extensive coevolution of insects and flowering plants, or that early biologists wrestled with a woefully incomplete fossil record, as explanations of or solutions to the abominable mystery (Friedman, 2009). As with any theory in biology, one approach cannot explain everything, and allowing for a close examination

of the peculiar genetics and genomic structure of angiosperms could be a crucial piece of this puzzle. Whatever the case may be for how these plants came to take the globe by storm upon their divergence some 200 million years ago or longer, the assortment and range of flowering plants have sparked endless inquiry for generations of scientists (Zeng et al., 2014). While a more complete fossil record and a more thorough understanding of evolution and its constraints has helped to clarify the origin and early history of flowering plants, much still remains to be revealed as to how plants came to be so wide spread and occupy so many niches.

The frequency of polyploidy, throughout time and geographical space among flowering plants, may help shed light on this abominable mystery and help to explain how angiosperms have come to define the diversity of life on earth. Other multicellular eukaryotes like animals and fungi tolerate polyploidy less frequently than plants, although the role of polyploidy and duplication have likely been underappreciated in these other eukaryotes (Mable, 2004). Regardless, polyploidy has played a crucial role in the evolution of flowering plants. Much like when Ohno championed the idea in 1970 that duplication was an important evolutionary driving force in fish and vertebrates, where the importance of these ideas were not appreciated until much later, it was not until more recently that polyploidy in plants was appreciated as anything more than a supposed "evolutionary dead end" (Stebbins, 1971). In fact, with the frequency of ancient polyploidy events found in plant histories, it is evident that at the very least, polyploidy has a neutral effect on evolution, or that it drives or contributes speciation and diversification in plants. As genetic resources for plants are growing at an impressive rate, there is no better time to investigate how the frequency of polyploidy and duplication in the evolutionary history of plants has shaped their genomes and evolutionary trajectories. Much work remains to be done to better understand the rules governing the role of polyploidy events affecting gene expression,

selective pressures, or adaptation in plants. For crop plants in particular, little is known of how polyploidy and duplicated genes have contributed to the domestication and improvement of these plants. A fuller understanding of the processes of polyploidy and duplication in crop genomes will contribute to meeting the challenges of increasing food production drastically to meet the needs of a growing world population using increasingly marginal land.

Legumes, encompassing many economically important crop species along with several model species, are a prime candidate for studying genome evolution via polyploidy or duplication. Learning more about how evolution by duplication has affected genomes throughout history can point to better ways to improve the genetics of crop legumes today. An appreciation of the history of the genomes humans use in breeding is crucial to understanding how they might be changed in the future. For example, without the knowledge of the diploid progenitors of peanut, the presence of tetrasomic recombination between the tetraploid peanut subgenomes and the possibility of introgression of alleles from these progenitors would not have been possible. Without the knowledge that soybean's genome has been duplicated many times in its past, the discovery and understanding of multiple genes controlling traits like flowering time would have been far more difficult. Without intimate knowledge of the intricacies of autotetraploid breeding in alfalfa, the yield gains seen in that feed crop in the last century would have been impossible. Thus, knowledge of the history of a genome is informative for directing its future.

As sequencing technologies continue to improve, our ability to understand and appreciate the ways genomes evolve and change also improves. Evidence is mounting that the structural diversity of plant genomes is far greater than was previously assumed; in general, most genotypes of plants studied do not share the same gene set, often due to the deletion or expansion

of certain genes or gene families among genetic lineages (Hirsch et al., 2014; Li et al., 2014). With the publication of genome sequences for the first few diverse, non-reference genotypes in crops, such as the maize PH207 genome (where B73 is the reference genotype) (Hirsch et al., 2016), a fuller picture of the role and extent of duplication and genome reorganization in genomes of crop plants is finally accessible and promises new frontiers for the fields of plant breeding and genetics. As more of these diverse non-reference genotypes within species are sequenced and assembled, the true breadth of structural and functional diversity in plants will come in to clearer focus, enabling new discoveries for geneticists and breeders alike.

Soon, rapid and economical genome assemblies will allow plant geneticists to assemble *de novo* the genomes of all or many of their breeding lines, in order to quickly and precisely understand the structural variation within breeding lines/populations, including which genes therein are prime targets for selection based on presence, absence, or other structural variation among genotypes. Whereas plant breeding in the past has been predicated on the assumption that individuals within a species share genic content but vary primarily in their alleles, a growing understanding of these newer structural variants will allow for a wider toolkit for selection and an expanded vocabulary of diversity of traits for crop plants. Traits that are controlled by genes with copy number variation or presence-absence variation are of particular interest to breeders, especially as they can complicate marker development and use, and in the past would have been entirely ignored. Furthermore, engineering traits under the control of genes with multiple copies or copy number/presence-absence variation requires knowledge of how gene dosage and copy number affect these traits. As more genomes and genotypes are sequenced and assembled in the coming years, then, the understanding of how duplication affects plant genomes and how it can

be leveraged via breeding and genome engineering will be of special interest to forward-thinking scientists and breeders.

The challenges and obstacles to feeding a world of almost 10 billion people by 2050 are numerous; however, opportunities to accelerate the pace at which farmers, breeders, and scientists can rise to overcome them are also at hand. While agronomy, climatology, soil science, economics, and artificial intelligence technologies all play a vital role in the task set upon those who work to feed the world, breeding and genetics will be just as important as they have been in the past. To accelerate the rate at which crop plants are improved, a deeper understanding of the genomes being manipulated and reshaped to human benefit is essential. To that end, scientists who want to understand the genetics of the plants they work with to create improved plants would do well to appreciate the role that polyploidy and gene duplication play in these genomes. Although these are important forces in the evolution and improvement of plants, deeper insights will be found through additional genome sequences and functional analyses that will lead to predictive models of the how variation derived from gene and genome duplications can more effectively be used for crop improvement.

OBJECTIVES OF THIS WORK

The dissertation work presented here has four major goals. First, it aims to analyze how soybean's multiple whole-genome duplication event have affected the organization, structure, and evolution of the soybean genome - especially through subfunctionalization of ohnologs - using an analysis of syntenic genome segments and their characteristics. Next, it will compare how soybean's duplications and resultant genome organization compare to another paleotetraploid and economically critical crop, maize (*Zea mays*), using a novel algorithmic approach to define ancient subgenomes and compare their characteristics (e.g. expression and

methylation). It then attempts to determine the nature of the most recent paleopolyploidy in soybean and how it differs from other *Faboideae* with a phylogenetic gene family approach, utilizing both a tree comparison algorithm and a machine-learning method. Finally, it aims to describe how gene or whole-genome duplication and presence-absence variation have affected the domestication of soybean via human intervention from wild *Glycine soja* into elite *G. max* lines using a large resequencing panel.

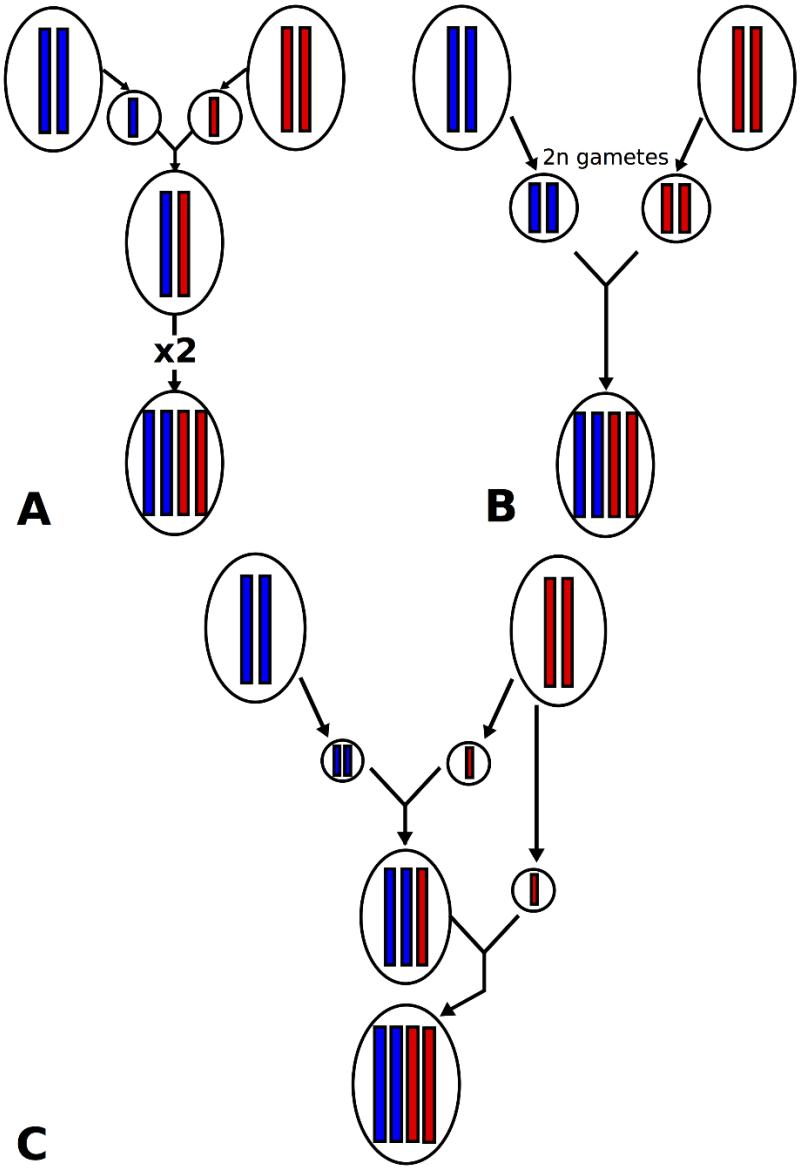

Figure 1.1. Schematic of different models of polyploid formation. A) a model typically shown in books, presentations, and other publications, but which is in reality quite rare and mostly not representative. B) A somewhat more common or accurate model, where 2n gametes are produced by both progenitors to create a tetraploid. C) The more likely unilateral model, where 2n gametes are produced by only one parent, and subsequent fertilization of the resultant triploid by the reduced gamete-producing parent creates a viable tetraploid.

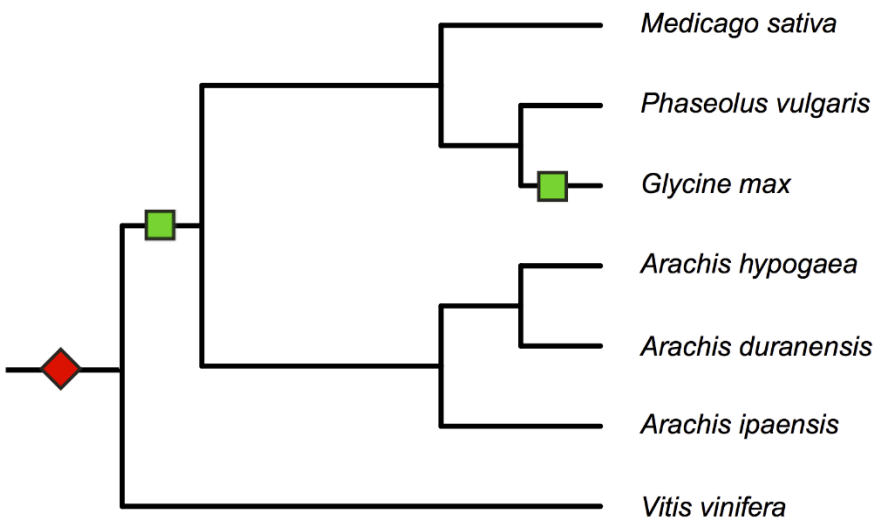

Figure 1.2. Simplified species tree of selected crop legumes alfalfa (*Medicago sativa*), common bean (*Phaseolus vulgaris*), soybean *(Glycine max*), modern cultivated peanut (*A. hypogaea*) and its diploid ancestors (*A. ipaensis* and *A duranensis*), and grape (*Vitis vinifera*) as an outgroup. The red diamond represents a hexaploidy at the base of the Core Eurosid clade about 125 Mya, the middle green square represents a tetraploidy ~60 My old shared by the *Faboideae* legumes, and the rightmost square represents the tetraploidy 8-13 My old specific to the *Glycine* genus.

CHAPTER 2

THE EVOLUTION AND DIVERGENCE OF WHOLE-GENOME AND SEGMENTAL DUPLICATIONS IN SOYBEAN

Introduction

Whole genome duplications (WGDs), both recent and ancient, are thought to be contributors to the evolution of life on earth, particularly in flowering plants (Flagel and Wendel, 2009; Gottlieb, 1982; Lynch and Conery, 2000; Ohno et al., 1968; Panchy et al., 2016). Although WGDs are instantaneous events, often with dramatic and immediately observable biological consequences, the long-term evolutionary consequences of gene duplication are impossible to observe directly. However, relics of WGDs can be observed in contemporary genomes as syntenic regions, collinearity, and an abundance of duplicated paralogs with clustered substitution rates (Casola and Lawing, 2018; Lynch and Conery, 2000; Tang et al., 2008a). Duplicated genomes experience various genome-wide changes over periods of time, and duplication often leads to speciation events (Schluter, 2001). On a whole-genome scale, polyploid genomes often revert to a diploid state via diploidization, which may be a relatively rapid process (Wolfe, 2001). On a smaller scale, duplicated segments or individual genes may be deleted, pseudogenized, or modified in the process of fractionation, which is often slow and takes place over long periods of time, often tens of millions of years (Freeling et al., 2015). While massive gene loss across a genome seems to be a catastrophic event *prima facie,* this process likely has some adaptive functions for a post-polyploid genome (Casola and Lawing,

2018; Gonzalo et al., 2018). The hypotheses and theories describing the evolution of these duplicated genes and segments are crucial for understanding how genomes change post-duplication, since experimental and direct observation of these extremely slow processes is not possible. As such, the "X-functionalization" hypotheses, the dosage balance or gene balance hypothesis, and dosage compensation hypothesis are useful for modeling or studying duplicate gene and genome evolution (Birchler and Veitia, 2012; Freeling et al., 2015; Lynch and Conery, 2000; Ohno et al., 1968; Panchy et al., 2016).

These X-functionalization hypotheses derive their name from the original neofunctionalizion hypothesis first coined decades ago, with the "X" representing one of several prefixes that predict the evolutionary fate of duplicate genes (Muller, 1936; Ohno, 1970; Ohno et al., 1968). The lack of widely available high-quality sequence data, or any whole-genome sequences at all, at the time of the conception of the neofunctionalization hypothesis severely hampered scrutiny, validation, or critique of this hypothesis, and researchers had to rely on sometimes difficult to interpret cytological or protein data (Muller, 1936). Nonetheless, the neofunctionalization hypothesis has attracted interest across disciplines and time in biology, and has evolved over time to include new observations and predictions, especially in light of data and observations generated by new sequence-based and molecular techniques (Flagel and Wendel, 2009; Freeling et al., 2015; Lynch and Conery, 2000).

Neofunctionalization predicts that after duplication, the two copies of a gene are functionally redundant; this means that loss or modification of one gene will often not adversely affect organismal fitness. Thus, one copy of the duplicated pair is free to accumulate mutations that alter the function of the gene, since it is under relaxed negative (purifying) selection and thus can potentially gain new functions (Assis and Bachtrog, 2013). Evidence of neofunctionalizion

has been observed across the tree of life, as in animals (Assis and Bachtrog, 2013; Logeman et al., 2017), plants (Emms et al., 2016; Gottlieb, 1982), and yeast (Byrne and Wolfe, 2007). As a corollary and later addendum to this hypothesis, subfunctionalization predicts that in some cases, the two copies of a duplicated gene might not gain new functions at all, but rather assume subsets of the original function of the non-duplicated gene (Panchy et al., 2016; Rastogi and Liberles, 2005). For instance, an ancestral gene with functions AB may duplicate and split its functions between the duplicates such that one gene takes on solely function A and the other takes on function B. Finally, nonfunctionalization predicts that in cases where gene function is unconstrained by dosage effects, one duplicate copy may lose all function altogether and become pseudogenized (Force et al., 1999; Lynch and Conery, 2000). Together, these X-functionalization hypotheses explain the scenarios in which duplicate genes diverge in their evolutionary history or in their biological function, and examples abound in the literature of duplicate gene families that have experienced one or several of these functionalization scenarios (Assis and Bachtrog, 2013; Blanc and Wolfe, 2004; Casola and Lawing, 2018; Cliften et al., 2006; Gottlieb, 1982; Kim et al., 2012).

Not all duplicated genes experience substantial divergence in their functions or histories, however. In some cases, duplicate genes maintain similar or identical functions over thousands to millions of years and are thus retained in duplicate with similar functions. For these genes, a different hypothesis is needed to explain why they have evaded functional divergence. The gene balance or dosage balance hypothesis represents an attempt to explain how certain genes appear to be retained in duplicate over long periods of time whereas many other genes are not (Birchler and Veitia, 2007; Birchler and Veitia, 2012; Veitia, 2004). It posits that genes that participate in regulatory networks, signaling pathways, multimeric complexes, or other multi-member

interactions are sensitive to changes in dosage in the networks they participate in (Veitia, 2004). Thus, these genes tend to be retained in duplicate because they might interact with a network, pathway or function that is mediated by a delicate stoichiometric balance. For instance, if proteins A and B interact in a 1:1 ratio to produce an important metabolite C, an interruption to this 1:1 ratio would upset the balance of the network and perhaps have a catastrophic impact on organismal fitness. If genes A and B were duplicated so there were now 2 copies of each, doubling the dosage of each member, the 1:1 ratio would still be maintained (now a 2:2 ratio); but if one copy of gene A were deleted or pseudogenized, the formerly healthy 1:1 ratio would become 1:2, and production of C might be impeded. Thus, the gene balance hypothesis predicts that genes with many network interactions, genes that depend on multimer formation, and genes that participate in regulatory or signaling pathways would tend to be retained in duplicate after WGDs (Teichmann and Veitia, 2004). Indeed, this is borne out by the observation that genes like e.g. transcription factors are often members of highly duplicated gene families retained post-WGD (Birchler and Veitia, 2007; Shiu et al., 2005; Xu et al., 2018).

Though WGD is recognized as a crucial force in genome evolution, there are many additional modes by which genes are duplicated. For instance, genes can be tandemly duplicated (i.e. duplicated adjacent to or very near their parent gene) through replication slippage, unequal crossover, or other mechanisms. They can also be duplicated in a dispersed fashion (far away from their parent gene, often on different chromosomes) via illegitimate recombination (or sometimes unequal crossover as well) (Bowman and Kurosky, 1982; Jelesko et al., 1999; Leister, 2004; Maere et al., 2005; Ohno et al., 1968). The evolutionary history of the events by which these tandem, dispersed, or other duplications transpire is far more difficult to untangle and interpret than that of WGDs, however, as these can occur continuously throughout organismal

evolution while WGDs happen in discrete, singular events. As such, any attempt to track the specific history of any pair or family or duplicated genes is made simpler by focusing on segmental or WGD duplication events, which can be deduced from analysis of synteny within and between genomes (Tang et al., 2008a). Many genomes, especially those of plants, have been analyzed for WGD and segmental duplications in this way starting with the publication of some of the first comparative maps of related plants (Bonierbale et al., 1988; Devos, 2005; Prince et al., 1993; Tanksley et al., 1988).

WGDs are instantaneous, discrete events that are time-estimable, and have occurred many times in the history of all characterized flowering plants. In the economically critical plant soybean (*Glycine max* L.), there are at least three WGD events that are detectable through synteny analysis: one shared with all core eurosids (~125 Mya), one shared with all papilionoid legumes (*Faboideae*, ~60 Mya), and another exclusive to the *Glycine* genus suspected to be concurrent with its first appearance (~8-13 Mya) (Lackey, 1980; Shoemaker et al., 2006; Vanneste et al., 2014b; Zheng et al., 2013). Each of these three events represents an opportunity to study the divergence of duplicate gene copies at three different stages of gene evolution: the very ancient (gene copies arising from perhaps the Jurassic or Cretaceous period), the ancient (duplicates from around the start of the Cenozoic era), and the relatively recent (from about the Neogene period). Furthermore, these WGDs arose from different biological, environmental, and cytogenetic constraints and are thus worth considering as unique events where genes evolving after each event were under different evolutionary pressures. For instance, some new evidence suggests that while the core eurosid hexaploidy (~125Mya) and the *Faboideae* duplication (~60 Mya) were likely allopolyploidies (arising from two different species' genomes coming together in one nucleus), the *Glycine* duplication event (~10 Mya) was perhaps more like an

autopolyploidy or segmental allopolyploidy (Wang et al., 2017a). This means that the initial states for these WGDs were vastly different, and thus these genes were also likely under very different constraints in their evolutions.

Like many legumes, soybean represents both a valuable economic product and an important model species. Furthermore, soybean appears to have an unusually high number of duplicated genes despite millions of years passing since its most recent WGD (Schmutz et al., 2010a). This, combined with its other two detectable WGD events, makes soybean a prime candidate for studying how repeated and varied WGDs can affect gene evolution under various contexts. In this study, various characteristics of syntenic blocks within the soybean genome are examined to discover how the differing ages and types of duplications found within a single genome have affected the evolution of duplicated genes in one of the most important crops on earth.

In this work, we created syntenic alignments of the soybean genome to identify WGD blocks and their age of duplication, revealing 3 major WGD clusters corresponding to at least 3 WGD events. We also found that expression of WGD gene copies is strongly linked in newer duplications, especially in the youngest *Glycine* specific duplication, suggesting subfunctionalization is highly reduced in these copies. Lastly, we found that methylation status, defined as one of four classifications, is potentially affected by or related to subfunctionalization, and that pairs of genes from the *Glycine* duplication event were potentially slowed in their transition from both being gene body methylated to one copy becoming unmethylated and the other maintaining gene body methylation.

Materials and Methods

The most recent version of the soybean annotation (Wm82.a2.v1) was downloaded from Phytozome 12. Full-length primary transcript-only CDS sequences for each gene were obtained from this annotation. A BLAST database was constructed with "makeblastdb" from these CDS sequences, and the CDS sequences were searched against themselves with an all-by-all BLASTn with an e-value cutoff of 1e-10. Using MCScanX, a set of syntenic blocks describing duplicated genome segments was built with default parameters (minimum gene number to call a block of 5, maximum gaps 25, gap penalty -1, match score 50, expect value cutoff of 1e-05).

Each alignment was a paired set of blocks of genes, and thus the sequence distance between each pair could be determined using the synonymous substitution rate per synonymous site. This was accomplished using the "add_ka_and_ks_to_synteny" function from MCScanX. Within each block (alignment) in the MCScanX output, the average Ks (synonymous substitution rate per synonymous site) was determined by adding up all the Ks values and dividing by the number of genes in the alignment. These block average values were then clustered using the "kmeans" function in R with k=3, given that there are 3 expected duplication events detectable in these soybean blocks (the *Glycine* WGD, the *Faboideae* event, and the core eurosid WGD). Each block was then assigned an age based on these clusters: 'Glycine', 'Papilionoid', or 'Ancient' (because the latter includes potentially some earlier, nearly undetectable events as well as some dispersed duplications).

To obtain estimates of expression for all these genes among 9 different tissues, a pre-existing and curated RNAseq dataset was obtained from Phytozome v10 (G. Stacey, unpublished data). The FPKM expression (fragments per kilobase of transcript per million mapped reads) for each gene was included in this dataset. The correlation between expression in each tissue for

duplicate gene pairs arising from each duplication age (27 total comparisons) was determined by plotting expression values of pairs of genes (transformed via $\log_2(1+FPKM)$) in a scatter plot and calculating the Pearson's correlation coefficient (R) for each of the 27 comparisons.

To investigate how methylation states may have changed over time between different ages of duplicated blocks, gene methylation state annotations were obtained from Niederhuth et al (2016). These assigned a state of unmethylated ("UM"), gene body methylated ("CG-gbm"), CHG methylated ("CHG-gene") or CHH methylated ("CHH-gene") based on a binomial test of uniquely-mapped bisulfite reads, a potential limitation given the similarity of recently duplicated genes in soybean. The numbers of genes in each category for all genes in the soybean genome, and all genes in at least one alignment belonging to each duplication age were counted. Furthermore, the paired methylation states for aligned pairs of genes from each duplication age were counted and compared, where every pairing was considered only if it was unique (i.e. an "UM/CG-gbm" pairing was considered the same as "CG-gbm/UM").

Results

Defining syntenic blocks and their ages in soybean

The first step in investigating how different ages and types of WGDs in a single genome have affected duplicate gene/genome evolution is defining those WGDs. In order to accomplish this, the complete CDS (coding gene sequence) for all genes and their coordinates were obtained from Phytozome 12 (genome Gmax 275 v2.0). These were used to generate a database of syntenic blocks using MCScanX with default parameters (minimum of 5 genes to call a syntenic block, expect-value threshold 1E-10) (Wang et al., 2012). This resulted in 1059 syntenic blocks, with a mean length of 31.9 genes, a median length of 10 genes, and a maximum block length of

1053 genes. Each syntenic block in this resulting table represented an alignment of a sequence of collinear genes aligned to one other sequence of genes elsewhere in the genome.

Using the MCScanX function "add_ka_and_ks_to_synteny", synonymous substitution rates, a common estimator of sequence divergence, were added to each pairwise gene alignment. The mean Ks (synonymous substitution rate) for each block was calculated, and the means of the Ks values for each block or alignment were clustered using k-means clustering with k=3, so that there were 3 possible values each mean Ks for a block could be assigned to: Ancient (syntenic alignments that correspond to the core eurosid triplication ~125 Mya, or other unidentified blocks), Papilionoid (the shared *Faboideae* WGD event ~60 Mya), and Glycine (~8-13 Mya, the most recent and *Glycine* specific WGD). This resulted in 318 'Ancient' alignments, 242 'Glycine' alignments, and 499 'Papilionoid' alignments. These alignments were comprised of 3597 (6.5% of all genes) total genes in 'Ancient' alignments, 15728 (28.3%) in 'Papilionoid', and 34325 (61.7%) in 'Glycine', indicating that newer duplications had more intact duplicate gene copies. In each age category, the number of unique genes with an alignment in that category varied drastically. Of 36,603 total WGD genes, 34,325 (93.8% of WGD duplicated genes) genes had a copy from the 'Glycine' event 8-13 Mya, 15,728 (43.0%) had a Papilionoid duplicate, and 3597 or 9.8% of WGD duplicated genes had an 'Ancient' duplicate. Figure 2.1 shows the distribution of the mean synonymous substitutions of the blocks clustered by estimated duplication age.

To investigate how blocks of syntenic genes degrade over evolutionary time as genes are deleted, pseudogenized, or otherwise lost, the numbers of genes within each alignment were calculated and associated with the duplication age assigned to each alignment. Figure 2.2 shows the distributions of the numbers of genes in blocks in each duplication age. In order, *Glycine*

blocks had more genes than Papilionoid blocks, which in turn had more than Ancient blocks (all significant below $p < 0.00021$, pairwise T-test with Holm correction). The *Glycine* blocks had some extreme outliers, 25 alignments with more than 200 genes, whereas there were only 2 such Papilionoid blocks and none in the Ancient blocks. The largest alignment, from the *Glycine* event, consisted of 1,053 pairs of genes from chromosomes 19 and 3.

Comparing expression between differently aged paralogs to evaluate subfunctionalization over time

Under the subfunctionalization hypothesis, it is expected that sequence divergence and functional divergence of two duplicated gene copies would be correlated (Blanc and Wolfe, 2004; Freeling et al., 2015). Specifically, as a pair of duplicated genes evolves over millions of years, it would be expected that as they accumulate mutations, are subjected to varying selective pressures, or experience genetic drift with age they will diverge in expression patterns (Flagel and Wendel, 2009; Freeling et al., 2015; Lynch and Conery, 2000; Rastogi and Liberles, 2005; Wang et al., 2017b). This would be apparent in differing expression levels in different tissues between pairs of paralogs. For instance, a highly diverged paralog pair could likely have high expression for one paralog in leaf tissue while the other copy has low expression in leaf tissue, with an inverse relationship in flower tissue. Furthermore, older duplicate pairs should have accumulated more mutations (or perhaps would have been deleted or nonfunctionalized), meaning older paralog pairs should show more divergent expression (Blanc and Wolfe, 2004; Libault et al., 2010; Panchy et al., 2016). It is worth noting, however, that expression in a given tissue does not capture the entirety of the functional profile of a gene product, and that other evidence like protein-protein interactions, cell localization, protein structure, or domain analysis

can offer a more complete picture of the function of a gene. For the purposes of this study, we restrict the analysis of function to expression profiling among varying tissues.

To investigate this, expression data for *G. max* was obtained from Phytozome, which included normalized expression data for 9 tissues: pod, root hair, leaf, root, nodule, seed, shoot apical meristem (SAM), stem, and flower. These expression values were originally represented in fragments per kilobase of transcript per million mapped reads (FPKM), but in order to facilitate interpretable linear comparisons, each expression value was transformed via $\log_2(1+\text{FPKM})$ (Rapaport et al., 2013; Zwiener et al., 2014). Since the synteny database created in the previous step consists of a series of pairwise alignments of blocks of genes in the *G. max* genome, each pairwise gene alignment was assigned an age based on the clustering of mean Ks values for each block. Then, the log-transformed expression value of each pair of genes in each of the 9 tissues was plotted as a scatterplot (Fig. 2.3). This resulted in 27 total comparison plots, and a Pearson's correlation coefficient ("r") was calculated for each tissue-age combination. This was done to evaluate whether gene expression between pairs of paralogs within a tissue correlates, or is associated, with age of duplication.

The correlation values for each comparison varied considerably, with a strong correlation of expression values among genes derived from the most recent ("*Glycine*") duplication in SAM tissue (r=.77372) and a weaker correlation in flower tissue for the oldest duplicates ("Ancient", r=0.29784). In general, the correlation of expression values within a tissue sample for duplicate genes was stronger for the most recent duplication event that gave rise to a paralog pair was. This is consistent with subfunctionalization, though it is important to note that these paralog pairs were defined by synteny. This means that any gene within a block that has lost its paralogous

sister gene elsewhere in the genome (through short deletion mechanisms or pseudogenization) is not represented in these graphs.

It was observed that many paralogs showed evidence of highly asymmetric expression in a particular tissue as seen in a preponderance of points (indicated by higher densities of points in the heatmap or a redder color) along the x and y axes in each plot (Fig. 2.3). This could mean that subfunctionalization was particularly strong among these paralog pairs, and that these genes had completely compartmentalized functions when compared to their paralogs. These genes are referred to as 'fully subfunctionalized' hereafter. These genes which were apparently fully subfunctionalized, having zero expression in one copy in a tissue but significant expression in the sister copy in that same tissue, are of interest, as they represent a set of genes that appear to be especially prone to subfunctionalization, and perhaps represent genes that do not tolerate redundancy in gene function. To investigate the kinds of genes that appear prone to subfunctionalization as defined this way, a GO term enrichment analysis was performed. First, genes that had zero expression (i.e. 0 FPKM) in a given tissue in one gene copy but at least some expression in the aligned paralogous copy in that same tissue within any syntenic block were extracted. Next, for each tissue, a Fisher's Exact Test was used to determine whether the proportion of genes having each given GO term (out of 1506 identified in total that were present in any syntenic block) was under- or over-represented in the given fully subfunctionalized gene-tissue combination. This resulted in a set of 48 unique GO terms in total that were found to be under or overrepresented in this set of fully subfunctionalized genes among the 9 tissues (Table 2.1). Among these, "nucleic acid binding" was universally under-represented, indicating that these gene types were apparently resistant to subfunctionalization. In contrast, terms like

"response to auxin stimulus" were overrepresented in all cases, indicating these genes were under strong subfunctionalization pressures in the tissues studied here.

Divergence in methylation status between paralog pairs from different duplication events

While expression is often used as an indicator of functional divergence of pairs of gene copies arising from ancient duplication events, there are perhaps other characteristics that mark gene pairs evolving divergently or convergently post-duplication. One often-discussed such characteristic is cytosine methylation, which is a state where cytosine residues in DNA sequences are modified with 5-methylcyotsine. While sequence divergence, gene deletion, and differential expression of gene copies can indicate subfunctionalization, nonfunctionalization, neofunctionalization, or gene balance effects at play among duplicated pairs, DNA methylation might also track with varying evolutionary pressures acting on duplicated genes (Kim et al., 2015). These sequence-level epigenetic modifications occur in several different contexts within plant genomes: CG, CHG, and CHH. CG and CHG are symmetrical and thus can be maintained through replication, but CHH is not and thus requires siRNA targeting to maintain CHH methylation (Bewick et al., 2016; Bewick and Schmitz, 2017; Tariq and Paszkowski, 2004).

While these three contexts of cytosine methylation can exist anywhere in the genome, there are general patterns (Zhang et al., 2018). CG methylation is often found within or near genes, as well as in repetitive elements or transposons. CHG methylation is often found in non-genic heterochromatin, but can occasionally be found in or near genes, especially in mutants of IBM1 (*increase in bonsai methylation 1*), which works to remove CHG methylation. CHH methylation is found most commonly in or near transposons and in CHH islands, and is generally maintained by RNA-directed DNA methylation (RdDM) (Bewick et al., 2016). In general, methylation in all three contexts is associated with reduced gene expression, but CG methylation

in the absence of the other two contexts is correlated with constitutive gene expression (Bewick et al., 2016; Bewick and Schmitz, 2017). Thus, when examining genic sequence, binning genes into categories that describe the general pattern of methylation in each gene can be an informative method for determining how CG methylation might be affecting genes (Niederhuth et al., 2016). Thus, annotations for each gene for the reference (Wm82) genotype from bisulfite data were obtained which classified each gene as unmethylated (hereafter "UM"), CG methylated (CG methylation and no CHG/CHH, hereafter "CG"), CHG methylated (CG and CHG methylated with no CHH, hereafter "CHG") and CHH methylated (having all three contexts, hereafter "CHG") (Niederhuth et al., 2016). The methylation status for each gene was compared across syntenic alignments in order to determine if methylation was associated with sequence and expression divergence between paralog pairs and duplication age.

Paralog pairs from each age of duplication (Ancient, Recent, Glycine) were binned together by age and methylation status. Fig. 2.4a shows the total number of genes included in alignments from each duplication age in each methylation category. Unmethylated genes dominated the landscape of these duplicated genes, with CG gene body methylated genes being the second most-common category. In each age class, the patterns of methylation were mostly similar, with UM genes being by far the most frequent followed by CG-gbm, with CHG and CHH genes being very uncommon (with CHH being interestingly slightly more common than CHG) (Fig 2.4b). Intriguingly, CG gene body methylated genes and CHG genes were almost equally frequent among all genes (Fig 2.4a), but CG-gbm genes were much more common than CHG genes among all duplication age classes (Fig 2.4b), indicating that these genes were perhaps less likely to be deleted among syntenic genes.

Because the expression of ohnolog pairs in different tissues or developmental stages can diverge over time as a result of subfunctionalization or neofunctionalization, it is also possible that methylation status can likewise evolve according to established models of duplication gene evolution. Indeed, some studies have indicated that characteristics like 24-nucleotide siRNA targeting (associated with CHH methylation accumulation) can play a role in subgenome differentiation post-WGD (Cheng et al., 2016). To test whether pairs of genes' methylation statuses are evolving convergently or divergently, each possible pairing of methylation statuses within alignments was also considered separately: UM-UM, UM-CG, CG-CG, etc., for a total of 10 methylation status comparison categories. Similar to the previous results, UM genes mostly dominated, with UM-UM being by far the most common pairing, and CG, CHG, and CHH genes being generally most commonly paired with UM genes (Fig. 2.3). However, while for nearly every pairing, ancient duplications had fewer genes than Papilionoid, which had fewer than Glycine in ascending order, CG-UM pairings were more common than CG-CG among Papilionoid and Ancient duplicated pairs than Glycine duplicated pairs, with a total difference of 228 more Papilionoid pairs. Surprisingly, despite an apparent trend of increasing gene counts with newer duplication ages, a two-way ANOVA indicates that only methylation pairing is significant as an explanatory factor for number of genes, with duplication age having an F-test p-value of 0.20259.

Discussion

Soybean's large, duplicated genome

That soybean's genic content consists primarily of duplicated genes has been known for many decades (Shoemaker et al., 2006), and results from this study corroborate this. While previous studies have noted via synonymous substitution rates or comparative mapping that 2 or

3 distinct ancient WGD events were detectable in soybean's genome, this study demonstrates that *a priori* defining WGD duplicated genes, excluding other genes from analysis (as these could have arisen from tandem duplication, transposition, or other mechanisms and not WGD), and analyzing the mean Ks or synonymous substitution rate within each alignment gives the clearest picture yet of the WGD events found in soybean's genome (Figure 2.1). A few syntenic alignments with Ks values of 2.0 and above appear to be either members of alignments that are soon to diverge too far to be recognizable, or duplicate blocks perhaps arising from an older WGD like the WGD shared by all flowering plants ~192 Mya (Murat et al., 2017). In general, newer WGD blocks had more genes in their alignments and more alignments overall (Figure 2.2). The Pearson's correlation coefficient for synonymous substitution rate and number of genes in a block was R=-0.2696, indicating a weak correlation among the two. More importantly, however, the differences between blocks of each age were significant among all comparisons as per pairwise T-tests, indicating that Glycine-specific blocks had the most genes, followed by Papilionoid and Ancient blocks in order. This indicates that, over time, deletion, pseudogenization, recombination, or chromosomal changes like translocation, inversion, or deletion remove or displace genes in long syntenic blocks from their neighbors (Tang et al., 2008a; Woodhouse et al., 2010).

There is evidence that chromosomal rearrangements were quite common in soybean and common bean, a close relative of soybean that lacks the Glycine duplication event (Hougaard et al., 2008). Deletions, translocations, and other chromosomal changes would disrupt the kinds of extremely long syntenic blocks found in the Glycine duplication blocks identified in this study (Edger et al., 2018). This suggests that many of these rearrangements occurred before the Glycine duplication event, or that they occurred in common bean and not soybean. A

comparison of synteny between soybean, common bean, and a third legume species like *Medicago truncatula* may help to corroborate either of these possibilities. Other studies have, for example, demonstrated that genome rearrangements are common among legumes like *Lotus japonicus*, *Medicago*, common bean, and peanut (Hougaard et al., 2008), which complicates imputation of ancestral chromosome states and the untangling of the complex cytogenetic history of the *Faboideae*.

36,603 out of 55,589 total genes in the soybean genome (65.85%) were placed into at least one syntenic alignment, indicating that the majority of genes in soybean have maintained some sort of duplicate copy from a WGD event, though many also have copies arising from other stochastic duplications like tandem duplications, dispersed duplications, and transposition events (Xu et al., 2018). As expected, newer duplication events retained more duplicate copies. That about 35% of genes did not have an identifiable WGD duplicate indicates that even though soybean has a large genome with almost twice as many genes as many of its diploid legume relatives, many of soybean's duplicate genes have nonetheless been deleted, pseudogenized, or diverged beyond paralogous recognition. Still, it appears that soybean's duplicated blocks are well-maintained, especially for a genome whose most recent paleopolyploidy was up to 13 Mya.

Expression divergence among WGD paralogs and their subfunctionalization over time

The evolution of duplicate genes and the canonical patterns they tend to follow have been theorized, observed, and refined for decades. Prime amongst these is the subfunctionalization hypothesis, which posits that because two duplicate genes have redundant functions, selection can be relaxed, and each copy is free to accumulate otherwise deleterious mutations. Subfunctionalization could be said to occur when each of a pair of duplicated genes takes on a mutually exclusive subset of the original functions of the ancestral unduplicated gene, resulting

in divergent expression patterns for each copy. For example, one copy may be expressed only in flower tissue, while its sister paralog is expressed only in stem tissue. However, genes often vary significantly in their expression and do not typically have binary expression ("on/off") in this manner. We observed that among WGD duplicate pairs in soybean, expression of paralogs differs considerably within and between tissues (Fig. 2.3). The density of gene pairs plotted along the diagonal (indicated as yellow colors in the heatmaps) demonstrates that while many duplicate gene pairs have divergent expression within tissues, many still have near identical expression levels within a tissue. Furthermore, although many gene pairs in each tissue had slightly or moderately divergent expression, there were a particularly large number of genes with no expression in one copy in each tissue and duplication age combination, indicated by the density of points along the x and y axes of the plots in Figure 2.3. It is evident, then, that subfunctionalization is not only a slow, ongoing process, but that its effects lie along a gradient rather than being binary: some gene pairs are somewhat subfunctionalized, maintaining slight but significant differences in expression profile, while others have presumably lost one of their ancestral functions entirely and have assumed a smaller subset of functions.

The set of fully subfunctionalized genes within each tissue revealed many commonalities among the functional classes of genes that tended toward subfunctionalization (Table 2.1). For instance, "nucleic acid binding" and "protein binding" were underrepresented among subfunctionalized gene pairs in all 9 tissues. Nucleic acid binding genes, which are often transcription factors, enhancers, or repressors that contribute to signaling pathways that act in networks, have been noted before to be resistant to deletion and nonfunctionalization and can often be found in large gene families within plant and animal genomes. Their underrepresentation among genes with a loss of function in a tissue (or potentially a gain of

function in one copy, which is presumably less likely) indicates that these nucleic acid binding genes are less constrained by redundant functions or overlapping expression of two copies of a gene. Furthermore, protein binding genes are usually involved in creating protein multimers as parts of gene networks. That these classes of genes are apparently highly resistant to subfunctionalization is in concordance with the dosage balance hypothesis, which suggests that genes in networks or pathways, or which participate in or form multimer interactions, tend to be retained in duplicate or maintain redundant functions in order to preserve the proper balance of products and reagents in the network or pathway.

In contrast to these nucleic acid binding genes, "heme binding" was overrepresented among fully subfunctionalized genes in all tissues. Heme binding proteins include important genes like leghemoglobin that contribute to the critical process of nitrogen fixation in legumes like soybean. Other terms like "oxidation-reduction process" were also overrepresented in fully subfunctionalized genes in all tissues, and this term includes genes involved in e.g. the electron transport chains in plastids and mitochondria. These fully subfunctionalized genes likely interact with few other gene products, do not participate in regulatory networks, or simply tolerate changes in dosage or function of one copy of a duplicate pair. This would allow one member of the pair to accumulate mutations while the other maintains the required dosage level needed to carry out the protein's function, eventually leading to subfunctionalization. These genes appear to have specific enzymatic functions, as heme binding genes have functions like binding a metal ion to buffer the concentration of oxygen, and oxidation-reduction genes pass electrons from one molecule to another – neither of which necessarily involves multimer formation or activity regulation via a complex network, though for example many heme proteins do form multimers to function properly.

While many gene pairs examined here were seemingly prone to larger divergence in their expression, and by extension prone to divergence in their functions, a surprising number of genes remained similar in their expression in each tissue. Notably, the correlation between expression levels of a gene pair within a tissue decreased dramatically with the age of the duplication event separating the gene pair (Figs 2.3,2.6). Despite an estimated upper bound of 13 million years passing since the Glycine duplication event, the correlation of expression within each tissue for Glycine age gene pairs ranged from 0.64923 to 0.70479, a relatively strong correlation indicating that most of these genes have very similar expression patterns to their Glycine WGD paralogs. By contrast, gene pairs from the Ancient (core eurosid and older) event had expression correlations as low as 0.2358, indicating little correlation. With the middle Papilionoid duplication event dated at around 60 Mya and the Ancient event(s) dated at 125 Mya, approximately 50 to 60 million years separates each successive duplication event, yet the difference in the correlation coefficient between Ancient and Papilionoid pairs was on average 0.09534778, and the difference between Papilionoid and Glycine pairs was .2844 – a nearly threefold difference despite that the distance between the older two duplications was about 60-65 My and the difference between the newer two duplications was about 50 My (Fig 2.6). This implies that duplicate pairs from the Glycine WGD have diverged in their expression less per million years than pairs from the Papilionoid and Ancient events. This concords with mounting evidence that the Glycine duplication event retains more duplicates than expected given its age, and that perhaps this Glycine duplication event was more like a segmental allopolyploidy or autopolyploidy than a strict allopolyploidy. It is suspected that biased elimination of duplicate genes, and biased expression of duplicate genes arising from WGD events is strong and acts relatively quickly after polyploidy to establish genome dominance in paleopolyploids of an

allopolyploid nature, but not in autopolyploids or segmental allopolyploids. Thus, soybean may belong to this second class of WGD duplicated genomes, whose duplicate genes are less diverged and experience less biased elimination or expression differentiation.

Methylation in WGD paralogs is also associated with duplication age

Deletion and expression divergence (subfunctionalization/nonfunctionalization) are well documented as either drivers or consequences of genome fractionation post-polyploidy, but less is known about how methylation might be affected by or drive the process of fractionation in paleopolyploid genomes. Techniques and standards for analyzing methylation, and the general patterns for methylation and other epigenetic marks in genomes are still evolving, but some generalities have emerged. In plants, epigenetic marks like cytosine methylation tend to be stable through generations and between tissues but can sometimes vary between genotypes and between closely related species in both heterochromatin and euchromatin. In general, methylation is more abundant in heterochromatic regions, especially in the CHG and CHH contexts, whereas the CG context can also be found in some genes. In general, it could be said that cytosine methylation varies and evolves at rates in plants much like those of DNA. Thus, it is of interest to examine how cytosine methylation might evolved in the context of polyploidy, diploidization, and subsequent fractionation over millions of years. Though cytosine methylation can be present in three contexts, the genes of the soybean genome were categorized as unmethylated, CG gene body methylated, CHG methylated, or CHH methylated based on in order to make meaningful distinctions about methylation status.

Within the soybean genome, unmethylated genes dominated and CG gene body methylated genes were the second most-common class. Interestingly, however, CHH class genes were nearly as common among all soybean genes as CG class genes, but CHH genes were far

less common than CG genes among WGD genes (Fig 2.4). Thus, it seems that CHH genes are unlikely to be maintained in duplicate post-polyploidy or are perhaps very prone to loss. CHH-class methylation includes methylation in the gene body in all three contexts (CG, CHG, CHH), which is a trait often shared by transposons or deep heterochromatin. Furthermore, methylation in all three contexts in this manner is also correlated with a decrease in gene expression. Thus, these CHH-class genes may be near transposons or within heterochromatin (e.g. in the pericentromere), where they are subject to accumulation of mutations and pseudogenization, nonfunctionalization, or loss over time, as they are typically lowly expressed or not expressed at all and thus not subject to purifying selective pressure (Kim et al., 2015).

Because pairs of collinear genes define the syntenic blocks created by WGD, the paired combinations of each methylation status were examined to see if any methylation status pairing (e.g. UM-UM, UM-CG) was overrepresented or underrepresented among all alignments. As expected, given the preponderance of unmethylated genes, UM-UM was by far the most common pairing. The next most common pairing was UM-CG, indicating that unmethylated genes still dominated but that CG gene body methylated genes were more common than other types of methylation, consistent with previous observations (Bertioli et al., 2016; Bewick et al., 2016; Bewick and Schmitz, 2017). Among nearly all methylation status pairings, the *Glycine* duplication had the most pairs, followed by the Papilionoid duplication, with the Ancient duplication(s) having the fewest members. Unlike all other methylation status pairings, however, CG-UM genes had more Papilionoid and Ancient pairs than Glycine pairs than CG-CG (with 228 pairs more in Glycine CG-CG than Glycine CG-UM). This could be because of more extensive repatterning of methylation among gene pairs from the Papilionoid and Ancient WGD in soybean such that some duplicated genes which were previously unmethylated were CG gene

body methylated, or vice versa. Alternatively, this could mean that gene pairs from the more recent Glycine event have resisted divergence in methylation status, maintaining CG-CG pairs that would otherwise have become CG-UM. Currently, little is known about how gene body methylation is established or its putative evolutionary function, and some plants seem to have dispensed with any gene body methylation altogether to seemingly no deleterious effect (Bewick and Schmitz, 2017). It is also possible that the unexpectedly high frequency of CG-CG Glycine gene pairs could simply be due to random chance, but testing this hypothesis would require, for example, generating models of methylation evolution for duplicate genes and testing the observed methylation status of *G. max* genes against those models.

Overall, the higher number of CG-CG pairs than CG-UM pairs from the Glycine WGD could mean that, like the tight expression correlation and low observed duplicate deletion rates of the Glycine pairs, methylation patterns among Glycine WGD pairs (especially those involving CG gene body methylated genes) have maintained their ancestral methylation states more than those from the older Papilionoid or Ancient events. This could be due to dosage balance effects imposing more stringent purifying selection upon these Glycine gene pairs, or perhaps because of the functional characteristics of these CG-gbm genes. Little is currently known about how gene body methylated genes might become unmethylated over time. Gene body methylated genes are commonly housekeeping genes (Edger et al., 2018), and housekeeping genes have been observed to be less likely to be tandemly duplicated and more likely to be segmentally duplicated (Cannon et al., 2004). Furthermore, housekeeping genes are often single-copy genes or low-copy genes, and are generally thought to be under tight dosage balance constraints (De Smet et al., 2013). In cassava, it has been demonstrated that expression and gene body methylation levels are positively correlated for duplicate pairs of genes – that is, the duplicate

pairs of genes with the highest expression divergence also have the highest divergence in gene body methylation levels (Wang et al., 2015). Thus, it stands to reason that in a genome with a high level of segmental duplication like soybean, the duplicated blocks from the most recent and most well-represented WGD (the Glycine event) will have more retained pairs of CG-gbm or housekeeping genes in concordance with the dosage balance hypothesis.

Conclusions

Duplications have long been recognized as an important evolutionary force, and are perhaps a primary driver behind the adaptive success of angiosperms across the globe (De Bodt et al., 2005). Understanding how processes like duplication, polyploidy, diploidization, and fractionation have affected genomes in the past helps guide predictions and expectations for how they might look in the future. This is especially important for crops, as domestication, improvement, and ongoing breeding efforts all rely on a sound understanding of the underlying genetics and genomics of each crop (Williams, 1964). While soybean has millennia of breeding history behind it, an intimate understanding of its cytogenetics and genomics has only been reached in recent decades, and much still remains unknown (Lackey, 1980; Schmutz et al., 2010a; Shoemaker et al., 2006). The needs of a growing human population, which will require food grown on fewer acres and more marginal land, demand a rapid and decisive response from breeders and geneticists (Ceccarelli et al., 2010). A deeper understanding of the genome of this valuable crop will be indispensable for improving its genetics to meet the challenges of a changing globe.

Though soybean was known to have a large, highly duplicated genome, the publication of a reference genome and subsequent study of its genetic content revealed a clearer picture of how precisely soybean's genome had been affected by its detectible polyploidy events ~130, 60, and

13 Mya. It was initially thought that *Glycine* had arisen from a tetraploidy between two distinct species (allopolyploidy) (Gill et al., 2009; Schmutz et al., 2010a; Shoemaker et al., 2006), but newer evidence has cast doubt upon this hypothesis (Wang et al., 2017a; Zhao et al., 2017). The evidence put forth in this research, showing that not only are genes from the most recent paleopolyploidy in soybean well-conserved, but display less expression and methylation divergence than would be expected given the timing of its previous paleopolyploidies, further suggest that a reevaluation of the tetraploid origin of soybean may be in order. This may have implications for future breeding of the species, as while domestication and improvement have certainly fixed many beneficial alleles in genes related to domestication syndrome and other agronomic traits, there may be an untapped wealth of valuable alleles in duplicate paralogs of these fixed genes (Hamrick and Godt, 1996; Lam et al., 2010; Li et al., 2014; Valliyodan et al., 2016). Studying how these duplicate genes might have been fixed through domestication or improvement in one copy but left unmodified in another could give clues as to how certain genes might be targeted for future improvement in the crop. While all knowledge of a genome is valuable, soybean in particular represents a rich opportunity for geneticists and breeders to apply this knowledge to one of the most economically valuable crops in the world and take strides toward improving the crop.

Table 2.1. Pearson correlation coefficients (r) of expression of gene pairs from each duplication events in 9 tissues.

	Ancient	Papilionoid	Glycine
Pod	0.33053	0.42854	0.68453
Root hairs	0.29017	0.38825	0.68402
Leaves	0.28092	0.37339	0.68471
Roots	0.23588	0.34753	0.64923
Nodules	0.28211	0.36283	0.65663
Seed	0.34968	0.43045	0.69194
SAM	0.31159	0.4322	0.70479
Stem	0.33521	0.42449	0.698
Flower	0.27766	0.3642	0.65807
Average	0.29930556	0.39465333	0.67910222

Table 2.2. GO terms enriched or depleted among fully subfunctionalized genes in each tissue. A fully subfunctionalized gene has zero expression in one copy and significant expression in the sister copy. An odds ratio above 1 implies enrichment (overrepresented) and an odds ratio below 1 implies depletion (underrepresented).

Tissue	GO term	Description	Fisher P-value	Odds ratio
Flower	GO:0006979	response to oxidative stress	1.65E-06	2.330732527
Flower	GO:0009733	response to auxin stimulus	6.36E-17	3.832592251
Flower	GO:0020037	heme binding	7.76E-13	2.329582046
Flower	GO:0003676	nucleic acid binding	2.86E-13	0.237236779
Flower	GO:0016491	oxidoreductase activity	1.74E-07	1.671024058
Flower	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3.95E-08	2.293770463
Flower	GO:0006470	protein dephosphorylation	1.09E-05	0.147359993
Flower	GO:0005524	ATP binding	9.02E-09	0.619105887
Flower	GO:0008234	cysteine-type peptidase activity	2.71E-05	2.937508897
Flower	GO:0005506	iron ion binding	4.15E-07	2.118755427
Flower	GO:0005515	protein binding	5.45E-12	0.618366718
Flower	GO:0005634	nucleus	1.92E-10	0.382316787
Flower	GO:0055114	oxidation-reduction process	7.71E-09	1.526712669
Flower	GO:0008270	zinc ion binding	1.62E-05	0.600396726
Flower	GO:0016758	transferase activity, transferring hexosyl groups	2.67E-05	2.438707712
Flower	GO:0004601	peroxidase activity	5.91E-06	2.313198832
Leaves	GO:0006979	response to oxidative stress	2.24E-08	2.319094382
Leaves	GO:0004190	aspartic-type endopeptidase activity	1.72E-05	0.208815074
Leaves	GO:0009733	response to auxin stimulus	7.86E-14	3.021615503
Leaves	GO:0020037	heme binding	4.58E-15	2.228672113
Leaves	GO:0042545	cell wall modification	6.43E-11	3.503222345
Leaves	GO:0004857	enzyme inhibitor activity	2.47E-12	3.345302238
Leaves	GO:0003676	nucleic acid binding	8.49E-14	0.324227163
Leaves	GO:0015238	drug transmembrane transporter activity	5.23E-06	2.648428742
Leaves	GO:0005840	ribosome	2.98E-13	0.209444947
Leaves	GO:0016491	oxidoreductase activity	1.83E-10	1.699108559
Leaves	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3.09E-08	2.091915848

Leaves	GO:0005524	ATP binding	1.66E-10	0.641840403
Leaves	GO:0015297	antiporter activity	5.23E-06	2.648428742
Leaves	GO:0030599	pectinesterase activity	6.43E-11	3.503222345
Leaves	GO:0005506	iron ion binding	4.54E-07	1.917345747
Leaves	GO:0005515	protein binding	5.22E-13	0.659409786
Leaves	GO:0006855	drug transmembrane transport	5.23E-06	2.648428742
Leaves	GO:0055114	oxidation-reduction process	2.79E-14	1.595863039
Leaves	GO:0016788	hydrolase activity, acting on ester bonds	4.27E-09	2.561480967
Leaves	GO:0005618	cell wall	1.29E-08	2.564394857
Leaves	GO:0003735	structural constituent of ribosome	4.26E-13	0.228944619
Leaves	GO:0022857	transmembrane transporter activity	1.97E-07	2.409037941
Leaves	GO:0006412	translation	2.29E-12	0.252718367
Leaves	GO:0004601	peroxidase activity	3.12E-08	2.368053687
Nodules	GO:0009733	response to auxin stimulus	1.23E-12	3.148350026
Nodules	GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	1.26E-07	2.575514311
Nodules	GO:0020037	heme binding	7.17E-09	1.97343092
Nodules	GO:0042545	cell wall modification	3.17E-06	2.86945542
Nodules	GO:0004857	enzyme inhibitor activity	1.18E-07	2.877927812
Nodules	GO:0003676	nucleic acid binding	1.11E-10	0.342341932
Nodules	GO:0005840	ribosome	5.44E-09	0.268625048
Nodules	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3.29E-06	2.002039942
Nodules	GO:0005524	ATP binding	1.34E-12	0.566285406
Nodules	GO:0003824	catalytic activity	2.06E-06	0.438390745
Nodules	GO:0008234	cysteine-type peptidase activity	2.20E-06	3.105229679
Nodules	GO:0015743	malate transport	2.75E-05	4.440571216
Nodules	GO:0030599	pectinesterase activity	3.17E-06	2.86945542
Nodules	GO:0005506	iron ion binding	6.82E-06	1.90530342
Nodules	GO:0005515	protein binding	3.00E-17	0.567039237
Nodules	GO:0006508	proteolysis	2.22E-05	1.634278809
Nodules	GO:0010333	terpene synthase activity	9.39E-06	5.019924921
Nodules	GO:0055114	oxidation-reduction process	3.38E-06	1.391279862
Nodules	GO:0016788	hydrolase activity, acting on ester bonds	1.19E-09	2.831743219
Nodules	GO:0003735	structural constituent of ribosome	8.72E-09	0.293661706
Nodules	GO:0006412	translation	4.69E-08	0.324185787
Nodules	GO:0046983	protein dimerization activity	1.89E-10	2.06581924
Pod	GO:0006979	response to oxidative stress	3.26E-16	3.436753315
Pod	GO:0009733	response to auxin stimulus	5.77E-12	3.128913823
Pod	GO:0020037	heme binding	6.35E-24	2.929549312

Pod	GO:0042545	cell wall modification	2.97E-12	4.191892222
Pod	GO:0004857	enzyme inhibitor activity	7.29E-11	3.500511994
Pod	GO:0003676	nucleic acid binding	5.02E-20	0.129017571
Pod	GO:0005840	ribosome	2.93E-07	0.325344222
Pod	GO:0016491	oxidoreductase activity	2.30E-06	1.577711882
Pod	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	9.90E-11	2.520614903
Pod	GO:0005524	ATP binding	3.28E-09	0.621648337
Pod	GO:0008234	cysteine-type peptidase activity	1.37E-05	2.947652754
Pod	GO:0030599	pectinesterase activity	2.97E-12	4.191892222
Pod	GO:0005506	iron ion binding	1.38E-09	2.317520664
Pod	GO:0005515	protein binding	6.95E-13	0.616732341
Pod	GO:0005634	nucleus	3.58E-08	0.466268211
Pod	GO:0055114	oxidation-reduction process	9.87E-16	1.731008464
Pod	GO:0005618	cell wall	5.19E-08	2.728613998
Pod	GO:0003735	structural constituent of ribosome	5.60E-07	0.350118794
Pod	GO:0006412	translation	3.11E-07	0.343421296
Pod	GO:0046983	protein dimerization activity	1.07E-07	1.896394922
Pod	GO:0004601	peroxidase activity	5.15E-15	3.425890205
Root	GO:0009733	response to auxin stimulus	5.40E-14	3.151584906
Root	GO:0020037	heme binding	1.03E-05	1.673973774
Root	GO:0042545	cell wall modification	2.63E-05	2.565139629
Root	GO:0004857	enzyme inhibitor activity	1.47E-05	2.371239043
Root	GO:0003676	nucleic acid binding	8.47E-11	0.371418318
Root	GO:0005840	ribosome	1.07E-09	0.283017129
Root	GO:0016491	oxidoreductase activity	2.02E-07	1.588833539
Root	GO:0005524	ATP binding	5.02E-08	0.67546152
Root	GO:0008234	cysteine-type peptidase activity	5.20E-06	2.878411082
Root	GO:0030599	pectinesterase activity	2.63E-05	2.565139629
Root	GO:0005515	protein binding	1.06E-11	0.66326204
Root	GO:0055114	oxidation-reduction process	1.05E-07	1.42366503
Root	GO:0006886	intracellular protein transport	1.42E-05	0.204866062
Root	GO:0003735	structural constituent of ribosome	1.70E-09	0.302553454
Root	GO:0006412	translation	6.45E-09	0.328123818
Root	GO:0046983	protein dimerization activity	1.78E-05	1.642554697
Root hairs	GO:0009733	response to auxin stimulus	3.68E-11	3.057357246
Root hairs	GO:0020037	heme binding	1.11E-13	2.340857975
Root	GO:0003676	nucleic acid binding	4.46E-11	0.313903974

hairs				
Root hairs	GO:0005737	cytoplasm	5.48E-06	0.219071689
Root hairs	GO:0005840	ribosome	9.97E-07	0.34941757
Root hairs	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	2.68E-10	2.480509096
Root hairs	GO:0005524	ATP binding	8.29E-09	0.628444474
Root hairs	GO:0005525	GTP binding	9.26E-07	0.270479228
Root hairs	GO:0016829	lyase activity	1.39E-06	4.607573074
Root hairs	GO:0005506	iron ion binding	3.47E-09	2.281976095
Root hairs	GO:0005515	protein binding	1.03E-09	0.667708859
Root hairs	GO:0010333	terpene synthase activity	2.93E-07	6.092368799
Root hairs	GO:0055114	oxidation-reduction process	1.05E-11	1.610111119
Root hairs	GO:0016788	hydrolase activity, acting on ester bonds	1.37E-10	3.03319203
Root hairs	GO:0003735	structural constituent of ribosome	1.79E-06	0.373569432
Root hairs	GO:0006412	translation	7.13E-06	0.405176455
Root hairs	GO:0046983	protein dimerization activity	4.57E-10	2.075455293
Sam	GO:0006979	response to oxidative stress	1.06E-08	2.460420763
Sam	GO:0009733	response to auxin stimulus	1.06E-15	3.377547642
Sam	GO:0020037	heme binding	1.22E-20	2.612509454
Sam	GO:0042545	cell wall modification	4.99E-06	2.757605548
Sam	GO:0004857	enzyme inhibitor activity	1.52E-06	2.601698053
Sam	GO:0003676	nucleic acid binding	5.10E-19	0.187549179
Sam	GO:0015238	drug transmembrane transporter activity	4.67E-07	2.990638304
Sam	GO:0005840	ribosome	4.10E-08	0.33004982
Sam	GO:0016491	oxidoreductase activity	6.58E-06	1.51427196
Sam	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	5.49E-12	2.516567447
Sam	GO:0005524	ATP binding	1.85E-06	0.709187267
Sam	GO:0005525	GTP binding	1.38E-06	0.320230229

Sam	GO:0008234	cysteine-type peptidase activity	3.41E-07	3.19755479
Sam	GO:0015297	antiporter activity	4.67E-07	2.990638304
Sam	GO:0055085	transmembrane transport	2.31E-06	1.552533898
Sam	GO:0030599	pectinesterase activity	4.99E-06	2.757605548
Sam	GO:0005506	iron ion binding	1.52E-10	2.305134471
Sam	GO:0016021	integral to membrane	1.02E-05	1.398299311
Sam	GO:0016020	membrane	2.94E-06	1.416628757
Sam	GO:0005515	protein binding	2.62E-11	0.66338782
Sam	GO:0006855	drug transmembrane transport	4.67E-07	2.990638304
Sam	GO:0055114	oxidation-reduction process	8.80E-11	1.534417533
Sam	GO:0016788	hydrolase activity, acting on ester bonds	1.21E-06	2.328349557
Sam	GO:0003735	structural constituent of ribosome	5.73E-08	0.348928802
Sam	GO:0006412	translation	1.84E-07	0.374992184
Sam	GO:0046983	protein dimerization activity	3.46E-10	1.998637025
Sam	GO:0004601	peroxidase activity	5.22E-09	2.582345976
Seed	GO:0006979	response to oxidative stress	5.15E-10	2.526017368
Seed	GO:0020037	heme binding	3.13E-13	2.141987598
Seed	GO:0042545	cell wall modification	3.08E-09	3.243338749
Seed	GO:0004857	enzyme inhibitor activity	5.01E-08	2.761166743
Seed	GO:0003676	nucleic acid binding	3.14E-21	0.186055651
Seed	GO:0005840	ribosome	3.57E-08	0.354735261
Seed	GO:0016491	oxidoreductase activity	1.88E-07	1.569911707
Seed	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	1.22E-05	1.853176745
Seed	GO:0030247	polysaccharide binding	8.87E-06	3.468489851
Seed	GO:0005524	ATP binding	4.35E-10	0.645564602
Seed	GO:0005525	GTP binding	1.32E-07	0.304511097
Seed	GO:0030599	pectinesterase activity	3.08E-09	3.243338749
Seed	GO:0004650	polygalacturonase activity	9.52E-06	3.579318289
Seed	GO:0005515	protein binding	8.63E-08	0.737824922
Seed	GO:0055114	oxidation-reduction process	1.74E-13	1.582045741
Seed	GO:0003723	RNA binding	3.51E-07	0.325954706
Seed	GO:0006886	intracellular protein transport	2.97E-06	0.18963329
Seed	GO:0016788	hydrolase activity, acting on ester bonds	2.76E-09	2.596316582
Seed	GO:0005618	cell wall	4.85E-08	2.507060505
Seed	GO:0003735	structural constituent of ribosome	3.44E-08	0.368798925
Seed	GO:0006412	translation	1.42E-08	0.361744108
Seed	GO:0046983	protein dimerization activity	1.51E-05	1.621497009
Seed	GO:0003700	sequence-specific DNA binding transcription factor activity	6.26E-06	1.44169051

Seed	GO:0004601	peroxidase activity	6.95E-10	2.587765889
Seed	GO:0008017	microtubule binding	3.05E-05	0.23792223
Stem	GO:0006979	response to oxidative stress	9.39E-07	2.251586872
Stem	GO:0009733	response to auxin stimulus	9.21E-11	2.893008893
Stem	GO:0020037	heme binding	1.86E-17	2.491967911
Stem	GO:0042545	cell wall modification	5.51E-12	3.987837178
Stem	GO:0004857	enzyme inhibitor activity	4.93E-08	2.911132759
Stem	GO:0003676	nucleic acid binding	2.26E-13	0.282698829
Stem	GO:0005840	ribosome	1.58E-08	0.296157855
Stem	GO:0016491	oxidoreductase activity	5.12E-08	1.648508274
Stem	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	4.61E-11	2.473182811
Stem	GO:0005524	ATP binding	9.19E-11	0.605698146
Stem	GO:0008234	cysteine-type peptidase activity	2.82E-05	2.78646702
Stem	GO:0030599	pectinesterase activity	5.51E-12	3.987837178
Stem	GO:0005506	iron ion binding	9.79E-10	2.26979914
Stem	GO:0005515	protein binding	1.20E-14	0.606197073
Stem	GO:0005634	nucleus	1.72E-07	0.508108942
Stem	GO:0009058	biosynthetic process	1.34E-05	2.280801452
Stem	GO:0055114	oxidation-reduction process	2.25E-13	1.631643454
Stem	GO:0008270	zinc ion binding	2.21E-05	0.631688876
Stem	GO:0005618	cell wall	3.03E-10	2.963057247
Stem	GO:0003735	structural constituent of ribosome	2.37E-08	0.318697175
Stem	GO:0006412	translation	1.17E-07	0.347474863
Stem	GO:0046983	protein dimerization activity	5.00E-06	1.723257775
Stem	GO:0004601	peroxidase activity	1.15E-07	2.440106964

Table 2.3. Counts of WGD duplicated genes belonging to each methylation category or pairs of genes belonging to each category pairing. “UM” = unmethylated, “CG” = CG gene body methylated, “CHG” = CHG methylated, “CHH” = CHH methylated.

Methylation class	Ancient	Papilionoid	Glycine
CG-CG	112	758	1465
CG-CHG	9	65	62
CG-CHH	10	81	93
CG-UM	307	1473	1245
CHG-CHG	0	5	23
CHG-CHH	1	16	43
CHG-UM	38	235	264
CHH-CHH	1	6	88
CHH-UM	66	453	515
UM-UM	2213	10631	13478
CG	564	3178	4340
CHG	52	334	426
CHH	84	570	838
UM	4814	23364	28948

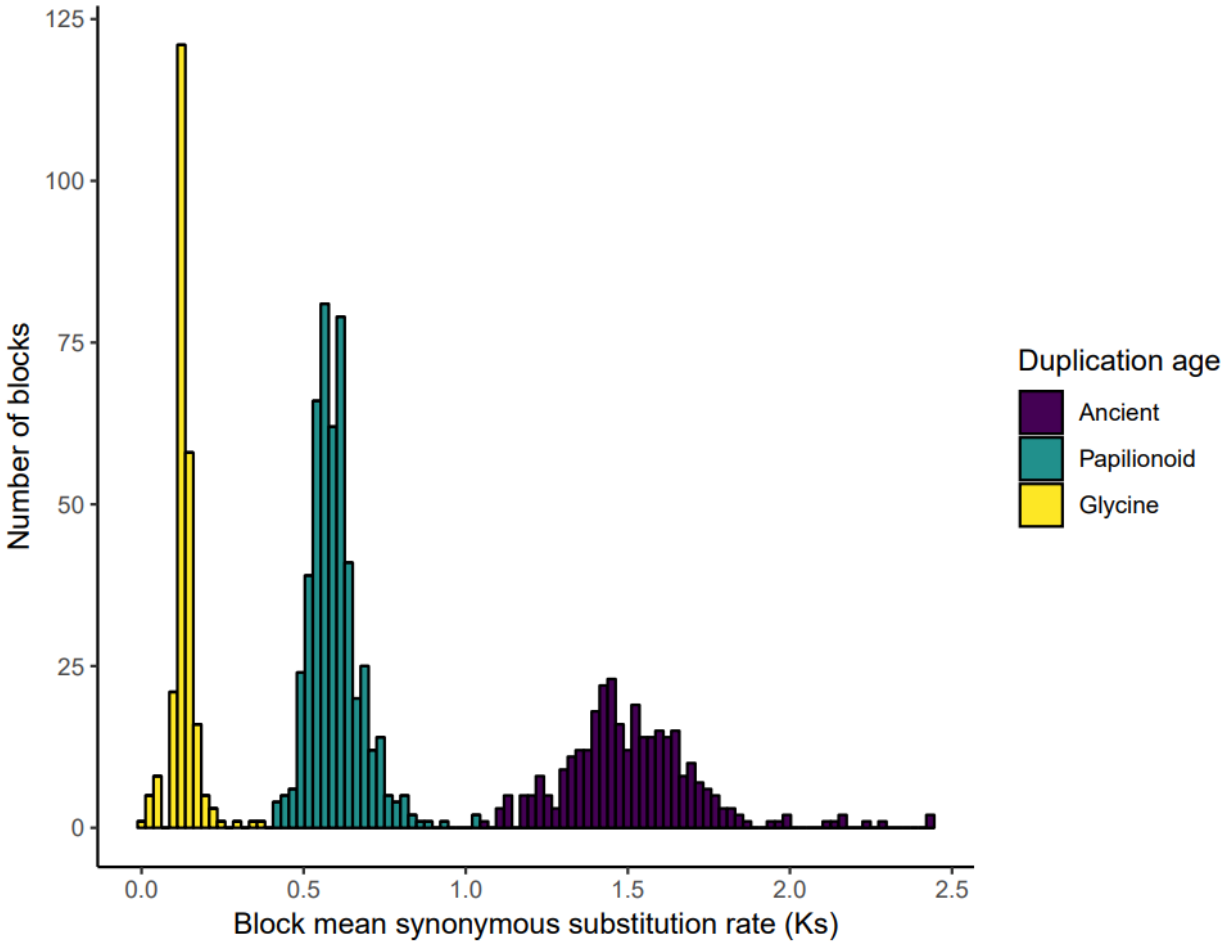

Figure 2.1. Distributions of mean synonymous substitutions per synonymous site rates (Ks) per syntenic alignment. Blocks were clustered using k-means clustering with k=3.

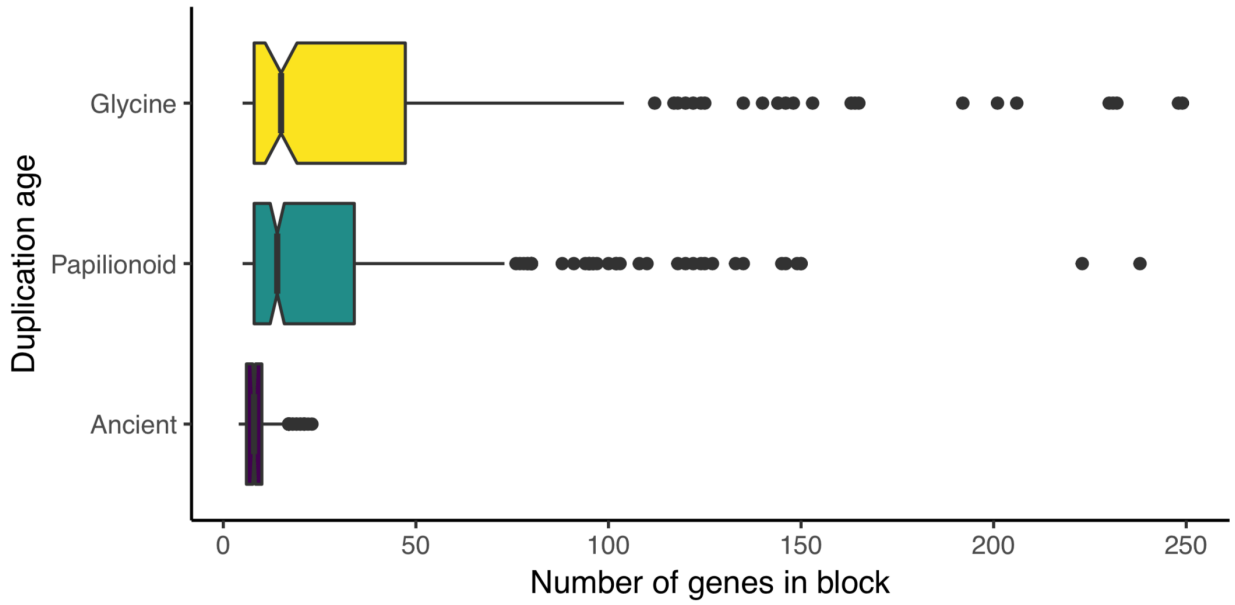

Figure 2.2 Boxplots of lengths of syntenic blocks by age. The number of genes in an alignment is used as a proxy for length. The mean length of the "Glycine" blocks was 71.39 genes, of "Papilionoid" blocks 27.5 genes, and of "Ancient" blocks 8.67 genes.

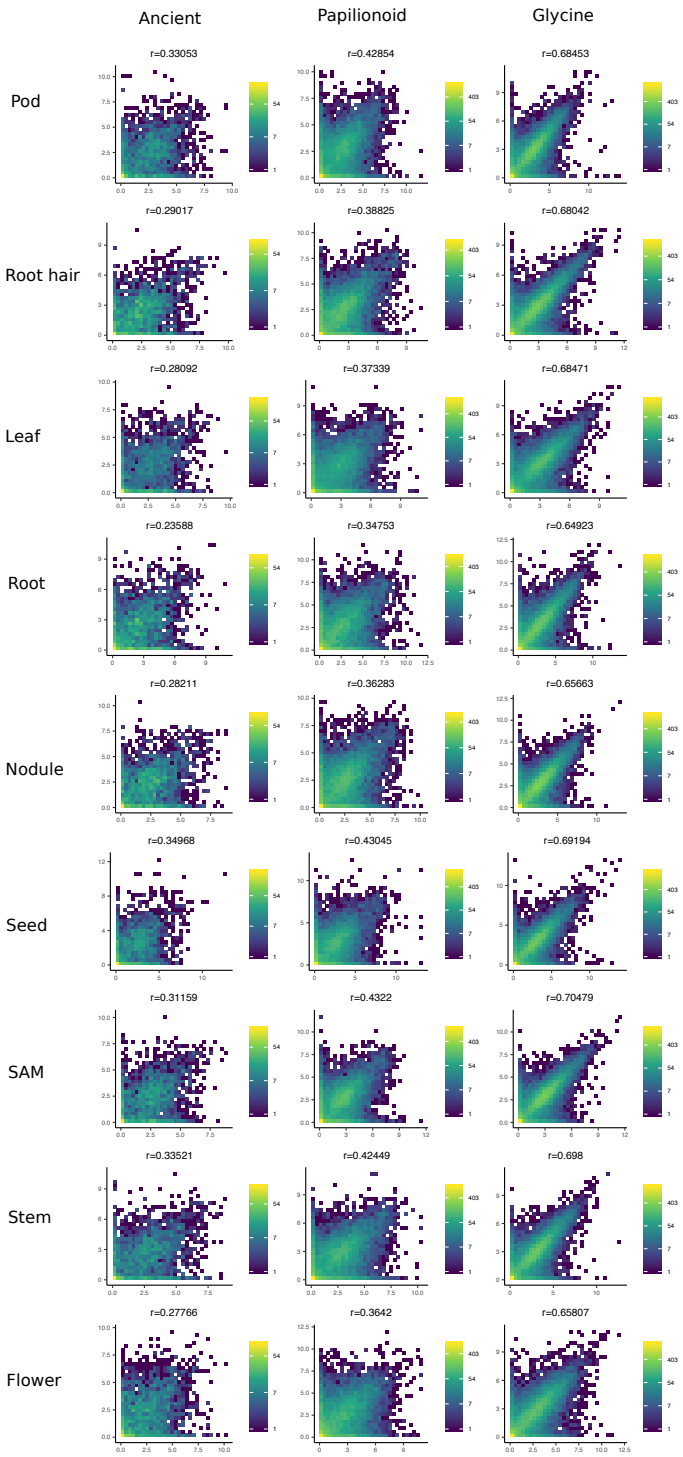

Figure 2.3. Correlation heatmaps of expression of WGD gene pairs in 9 tissues for each duplication age. The Pearson's Correlation Coefficient (r) is included at the top of each plot. The point density color scales in each bin on the right-hand side of each plot are in a $\log_{10}$ scale.

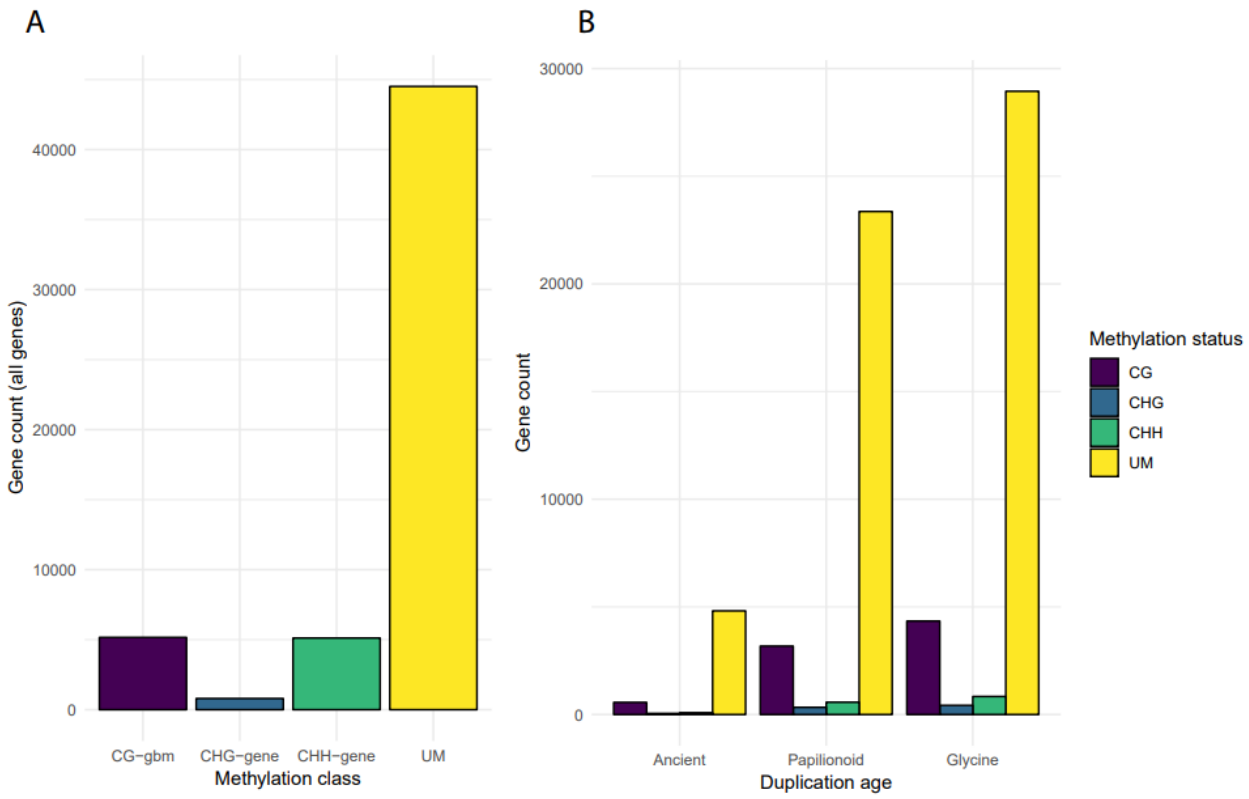

Figure 2.4. Proportions of genes belonging to each methylation class. A) Number of unmethylated (UM), CG gene body methylated (CG-gbm), CHG methylated (CHG-gene), and CHH methylated (CHH-gene) genes among all 55,518 soybean genes. B) Number of genes of each methylation class among only WGD duplicated/syntenically aligned genes, separated by duplication age. Methylation categories adapted from Niederhuth et al 2016.

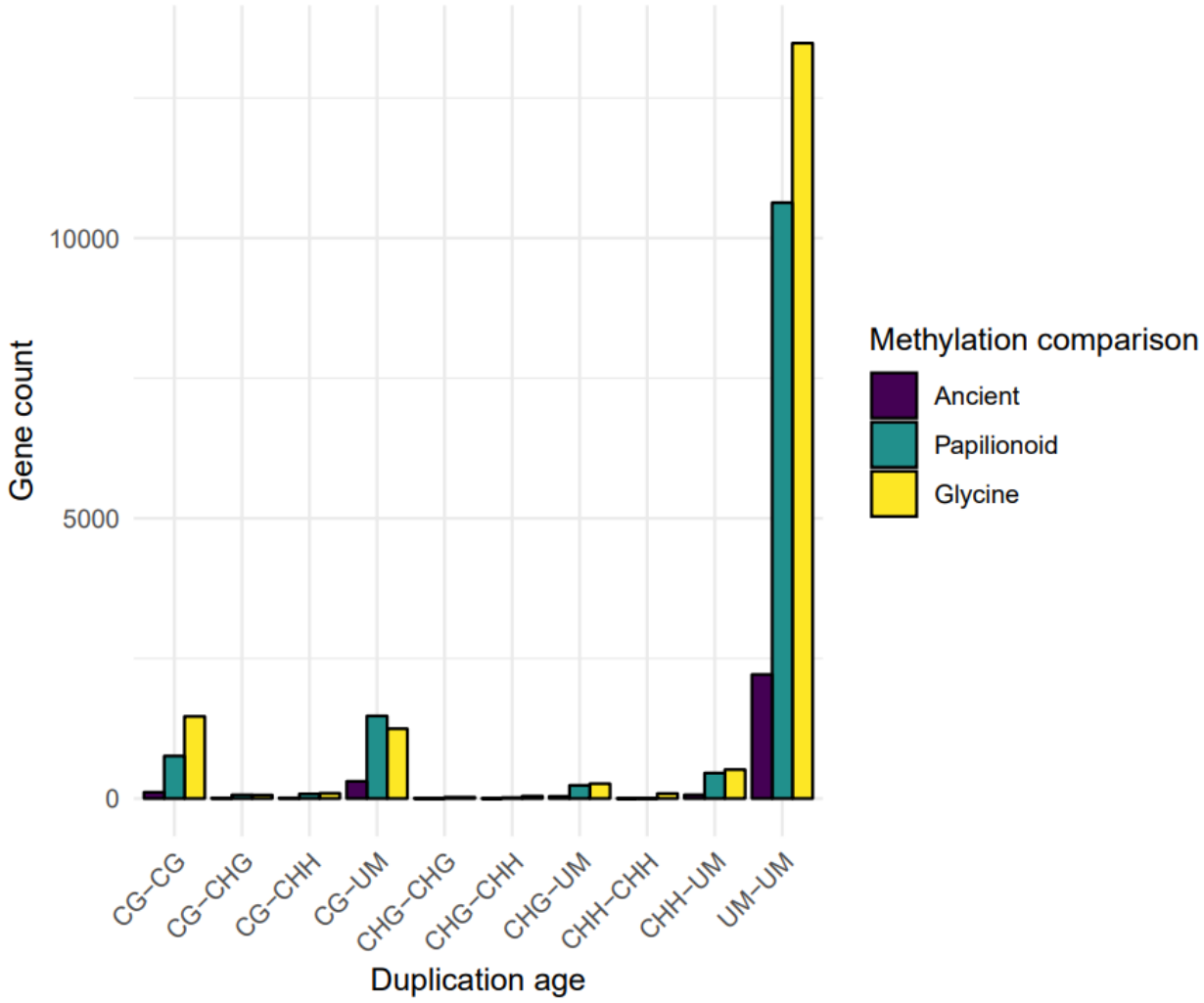

Figure 2.5. Number of syntenic gene pairs belonging to each methylation category, separated by duplication age.

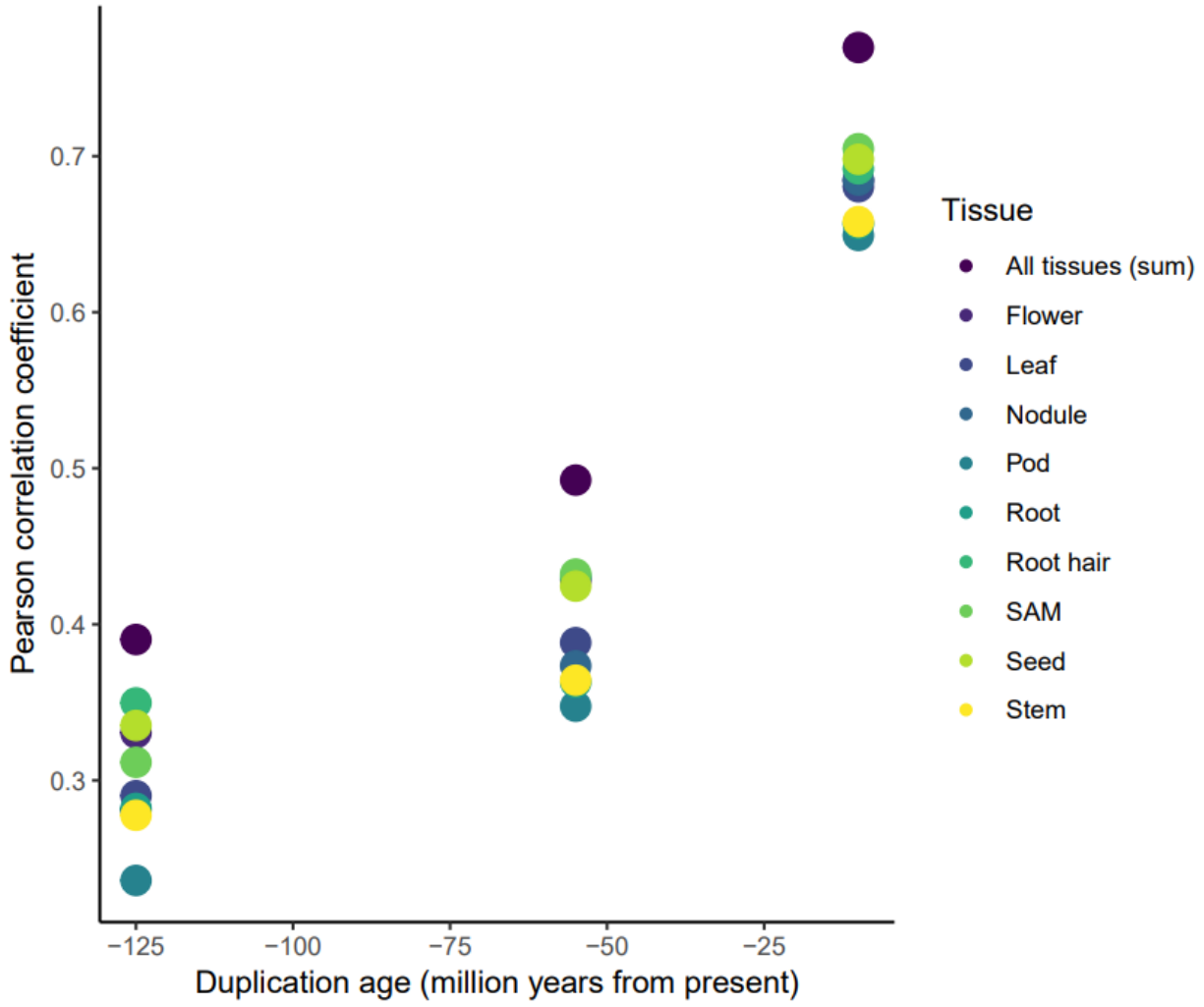

Figure 2.6. Correlation coefficients of expression in 9 tissues (plus the sum of expression across all tissues studied) for gene pairs arising from each duplication age. The X-axis is arranged negatively with respect to age (i.e. the oldest duplication is on the left).

CHAPTER 3

RECONSTRUCTION AND COMPARISON OF ANCIENT PALEOPOLYPLOID SUBGENOMES: CONTRASTING EVOLUTIONARY HISTORIES FOR SOYBEAN AND MAIZE

Introduction

All crop plants cultivated by humans today are either polyploid or have some history of ancient polyploidy (Jiao and Paterson, 2014; Murat et al., 2017; Paterson, 2005; Salse, 2016; Zheng et al., 2013). A doubling, tripling, or greater multiplication of the size of a genome in a single event, as in polyploidy, has dramatic effects on cell size, gene expression, reproduction, adaptation, and essentially every facet of the biology of an organism (Comai, 2005; Semon and Wolfe, 2007; Song et al., 2012). After a whole-genome duplication (WGD) event or polyploidy, however, there is a marked tendency for duplicated genomes to revert back to a diploid-like state over time, that is bivalent chromosome pairing at meiosis (Bowers et al., 2003; Van de Peer et al., 2009; Wolfe, 2001). Thus, WGDs not only precipitate sweeping biological changes of their own but are often followed by rearrangements of the genome in a return to a diploid state. These commonly include fragmentation or fusion of chromosomes, translocations or inversions of chromosome segments (Freeling et al., 2015), and deletions of duplicated chromosomes (Lin et al., 2010; Mandáková et al., 2010). The relics of these processes are present in all crop genomes and have implications for how these crops have evolved both in the wild and in human hands.

During the reorganization of a polyploid genome, duplicated genes and genome segments are often deleted, psuedogenized, silenced, translocated, or otherwise changed irreversibly

(Lynch and Force, 2000; Subramaniam et al., 2013). This general process of loss or rearrangement of duplicated genome segments is often called fractionation (Cheng et al., 2013; Garsmeur et al., 2014; Schnable et al., 2011; Tang et al., 2012; Wendel, 2015). Thus, while ancient duplications are often detectable by scanning for duplicated, syntenic genome segments where long runs of homologous genes are present in the same order in two or more places in a genome, this is often complicated by the process of fractionation (Haas et al., 2004; Keller and Feuillet, 2000; Proost et al., 2012; Tang et al., 2011; Wang et al., 2017a; Wang et al., 2012). Deletions of otherwise syntenic genes, and chromosomal inversions and translocations complicate detection of these blocks.

Ancient large-scale genomic changes like polyploidy, gene duplications, or chromosomal duplications and deletions have left indelible marks on modern genomes and thus have likely affected loci and genes that characterize important agronomic traits of the most valuable crops grown on earth (Comai, 2005; McClintock, 1993). Wheat, for example, shows considerable plasticity and adaptability which has been suggested to be a resulted of its hexaploid nature, which contributed to its domestication and widespread cultivation (Dubcovsky and Dvorak, 2007). QTL (quantitative trait loci) related to seed shattering have been mapped to duplicated regions on maize chromosomes 1 and 4, demonstrating that duplicated chromosome segments can be contribute to crop domestication (Buckler et al., 2001; Lin et al., 1995). It is critical, then, to understand these WGDs when examining how evolution and human selection has shaped wild plants found across the planet into the abundant, nutritional, and reliable staples of the human diet today.

Ancient WGDs (or paleopolyploidies) can broadly be categorized into two types, defined by the differential behavior of the subgenomes in the resultant polyploid. Class I WGD events

are thought to arise from polyploidies where the parent subgenomes are different from each other (e.g. allopolyploidy) and show marked differences in how these subgenomes evolved (Edger et al., 2017; Emery et al., 2018; Parkin et al., 2014; Wang et al., 2014), and Class II WGD events arise from polyploidies where the parent subgenomes are similar and show few structural differences (i.e. autopolyploidy) (Garsmeur et al., 2014). Importantly, these classes do not segregate neatly into binary categories, nor can they be definitively associated with either allopolyploidy or autopolyploidy. This is because not only is it often impossible to determine the exact or closest progenitors of an ancient polyploidy, but the auto- or allopolyploid nature of the event that gave rise to these polyploid genomes is often unclear (El Baidouri et al., 2017; Kochert et al., 1996; Marcussen et al., 2014). Instead, these are general classifications that exist on a spectrum, where a WGD event or genome can show characteristics of either or both classes of ancient WGDs. These classes of paleopolyploidy can be defined by the behavior of the parental genomes ("subgenomes" from here on) that contributed to the polyploidy event. In general, Class I events involve a pair (or more) of distinct subgenomes with marked differences in their duplicate gene deletion/retention, gene expression, methylation, and transposon content, while Class II events involve a set of (possibly indistinct) subgenomes with similar expression, methylation, and retention levels (Emery et al., 2018; Garsmeur et al., 2014). Notably, class I paleopolyploid genomes display a tendency for one subgenome to have fewer deleted genes, higher expression levels, and lower methylation levels than the other in a phenomenon often dubbed "genome dominance" (Chang et al., 2010; Edger et al., 2017; Flagel and Wendel, 2010; Freeling et al., 2012; Garsmeur et al., 2014). Class II genomes, on the other hand, have been observed to have little to no genome dominance and moreover are difficult to define subgenomes (D'Hont et al., 2012; Garsmeur et al., 2014; Wang et al., 2017a).

Previous evidence indicates that maize and soybean, respectively, are exemplars of Class I paleopolyploid and Class II paleopolyploid organisms on top of being two valuable crop plants worldwide (Garsmeur et al., 2014; Schnable et al., 2011; Wang et al., 2017a). Furthermore, these species both show a recent WGD with very similar timings: maize has a paleotetraploidy at about 10-12 Mya and soybean has a paleotetraploidy at about 8-13 Mya (Schmutz et al., 2010b; Schnable, 2015). This makes these two species an excellent set of species for comparing divergent evolutionary histories of two vital crop plants. Evidence from early sequence data on maize showed that there were distinct, observable differences in the gene expression levels and gene deletions between two subgenomes of maize arising from its most recent tetraploidy about 12 Mya (Freeling et al., 2012; Schnable, 2015; Schnable et al., 2011). These subgenomes were described by manually comparing syntenic segments of maize to a close relative, *Sorghum bicolor*, which lacked this tetraploidy. The results showed that, consistently, one 'subgenome' of maize had higher levels of gene deletion and lower expression levels, and furthermore that diverse maize lines had maintained this bias. In contrast, newer annotation and assembly data for soybean has shown little evidence of sub-genome bias, despite older data suggesting otherwise (Lin et al., 2010; Wang et al., 2017a). However, many disparate, incompatible, or difficult-to-reproduce methods have been used in the past to make these findings. Thus, comparing these two crop genomes with a set of consistent, algorithmic methods for identifying ancient subgenomes and ascertaining the presence of bias therein will elucidate how these two species with divergent evolutionary histories truly differ in their bias and post-WGD evolution.

In this study, a new algorithm for reconstruction of ancient tetraploid subgenomes (TetrAssign) was developed to solve the problem of determining subgenome states for highly rearranged paleopolyploid genomes like soybean. Applying the algorithm to both soybean and

maize revealed consistent biases in gene deletion rates, expression levels, and gene methylation status between maize's two ancient tetraploid subgenomes but not for soybean's. These results indicate that perhaps maize's ancient tetraploidy was more like an allopolyploidy or class I paleopolyploidy, and that soybean's was more like an autopolyploidy or segmental allopolyploidy or class II paleopolyploidy.

Materials and Methods

Full primary-transcript-only CDS gene annotations for *Glycine max* (Wm82 v2.0) and *Zea mays* (AGPv3) were obtained from Phytozome and Ensembl, respectively. Because each species arose from an ancient tetraploidy event, meaning that every gene should have 2 copies compared to its closest relative lacking the tetraploidy event, these two genomes were compared with related species lacking these more recent duplications: *Phaseolus vulgaris* (v2.1, Phytozome 12) for comparison to soybean and *Sorghum bicolor* (v3.1.1, Phytozome 12) for comparison to maize. Primary transcript-only CDS sequences were obtained and filtered for primary transcripts only for these species as well. Then, the *P. vulgaris* CDSs were used as a BLAST query (BLASTn 2.7.1) against *G. max*'s CDSs and the two best *G. max* CDS sequence matches for each *P. vulgaris* gene were retained (e-value cutoff 1e-10). The same procedure was performed using sorghum CDSs against maize sequences, retaining the two best maize hits per sorghum gene.

MCScanX (Wang et al., 2012) was used to chain the two best BLAST hits per 'reference' gene into groups of collinear gene blocks. A table was then created using these estimated collinear blocks, with the genes from the beginning to the end of each reference chromosome (i.e. common bean or sorghum) in the leftmost column, and genes matching (i.e. orthologous) to

the reference gene from maize or soybean in the subsequent right columns, with the initial goal of ensuring each reference gene has at most two orthologous genes from the corresponding paleotetraploid species. This was used as a starting point for inferring the ancestral genome state, pre-paleotetraploidy, for maize and soybean.

In order to resolve and estimate ancestral chromosome organization in maize and soybean, a new collinear block sorting and alignment algorithm was created and implemented in Python 2.7 (available at https://github.com/briannadon/TetrAssign). The basic rules of the algorithm set out to "phase" the blocks according to parsimony, and thus attempted to recreate ancestral chromosome order prior to the most recent tetraploidy event in soybean and maize by assuming the fewest chromosomal rearrangements and the fewest number of chromosome breakpoints possible. As such, blocks were sorted using this basic algorithm, using common bean/soybean as an example (Fig. 3.1):

1) Place all soybean blocks that align to a given common bean chromosome into an array (e.g. all soybean blocks that align to Pv01);

2) Find the first block from soybean that aligns to a common bean gene, and group all blocks from the soybean chromosome this first found block belongs to together into a putative 'subgenome' (for example, if the first soybean block that aligns to Pv01 is from Gm17, find all other blocks that align to Pv01 that are from Gm17 and group them together);

3) Search all the common bean chromosome genes that are covered by this soybean block, and find any collinear blocks that also cover this region that are from a different soybean chromosome, and assign those to the second 'subgenome' (e.g. there is a block from Gm14 that covers the same Pv01 region, so assign that to 'subgenome' 2)

4) Iteratively repeat steps 2) and 3) until all blocks in the array have been assigned to a 'subgenome'

After 'subgenomes' were assigned for both species, deleted genes (genes in the reference chromosome with no ortholog in either or both subgenomes), gene expression, and gene body methylation were examined across these reconstructed ancestral subgenomes.

To measure the number of deleted genes across each subgenomes assignment for each species, the total count of deleted genes across each reference chromosome was calculated using (common bean or sorghum) a sliding window of 100 genes with a step of 1. A Wilcoxon rank sum test with continuity correction was performed for each reference chromosome to compare the runs of gene deletions between subgenomes.

To compare expression between assignments, leaf expression data for *G. max* was obtained from young leaf tissue (AB036TABXX), while expression data for leaf tissue in maize was obtained from NCBI SRA reads SRR5368994. These reads were quality filtered and mapped to the maize genome (AGPv3) and the soybean genome (Wm82 v2.0) with STAR, filtering the GTF annotations to only include exons from primary transcripts. Transcripts per million (TPM) was calculated to quantify expression for each gene for each species. Each gene was then associated with its appropriate TPM. Then, similarly to the above protocol for comparing gene deletions, a sliding window average of expression levels was calculated by using a window of 100 genes on the reference chromosome with a step of 1. The average in each window was calculated as the average of the $log_2(1+TPM)$ for each gene in the window, with missing genes (i.e. deleted genes) ignored. A Wilcoxon rank sum test with continuity correction was performed for each reference chromosome to compare the overall expression bias between each subgenome. Pairs of genes between maize and soybean subgenomes were also compared by

classifying gene pairs as “different” or “similar” in their expression if either copy had >2-fold the expression of its sister copy, with a total of 12,707 soybean gene pairs and 3,201 maize gene pairs.

Lastly, to identify whether divergence in methylation states also tracked divergence in subgenome evolutionary history, binomially-assigned methylation state annotations for all genes in soybean and maize were obtained from Niederhuth et al (2016). Each gene could be assigned as “UM” (unmethylated), “CG-gbm” (CG gene body methylated), “CHG-gene” (CHG methylated), or “CHH-gene” (CHH methylated). Then, pairs of genes were classified as “different” or “similar” depending on whether both genes had the same or different methylation states.

Results

Identifying orthologs and paralogs to reconstruct syntenic blocks and ancient subgenomes

Both soybean and maize have WGDs of approximately the same age (5-13 My), but in order to determine what these genomes looked like before these recent WGD events, well-characterized and closely related species that lack these WGDs are necessary. These genomes must be assembled to a chromosome scale, as the order and orientation of large segments of genes is critical for trying to infer ancestral states. In soybean’s case, *Phaseolus vulgaris* serves as the point of comparison, as it has a well assembled and annotated genome and is a close legume relative that lacks the *Glycine* lineage-specific WGD found in soybean. Maize, on the other hand, has previously been compared to *Sorghum bicolor* for similar reasons, and thus *S. bicolor*’s genome will be used as the point of comparison for maize. Full coding sequence and gene coordinate annotations were downloaded from Phytozome.org (soybean, common bean, sorghum) and Gramene.org (maize) (Goodstein et al., 2012; Tello-Ruiz et al., 2018). Each

common bean gene is expected to match 2 soybean genes, and each sorghum gene is expected to match 2 maize genes owing to the WGDs in each lineage. Thus, an all-by-all BLASTn search (Camacho et al., 2009) was performed for each pair of species, where the best 2 soybean genes were matched per 1 common bean gene, and the best 2 maize genes were matched per 1 sorghum gene. These BLAST results were then clustered into syntenic blocks between the two pairs of species using MCScanX (Wang et al., 2012) with default parameters.

BLAST alignments and synteny scans resulted in 390 alignments generated for sorghum-maize and 604 alignments for soybean-common bean. In most cases, this resulted in 1 or 2 syntenic blocks from each paleotetraploid species aligned to each diploid species. In rare cases, however, 3 blocks aligned to the same region in a diploid species. In these cases, these blocks were collapsed to fit into one of the other blocks aligning to that region. In all, 62.47% of genes in soybean and common bean were included in these blocks, and 41.56% of maize and sorghum genes.

Reconstructing imputed ancestral subgenomes for soybean and maize with an algorithmic approach

In order to group syntenic blocks into a consistent, reproducible set of 2 ancient subgenomes per paleotetraploid species, an algorithmic, parsimonious approach was chosen. Previous studies have simply aligned orthologous genome segments and manually chosen blocks of genes to be included in differing subgenomes (Schnable et al., 2011; Wang et al., 2017a), and while this can result in usable subgenome assignments, the goal of this study was to use a consistent, reproducible method for comparing these assigned subgenomes between species. Thus, some cytological and evolutionary assumptions were made for the purpose of achieving this: 1) Intrachromosomal rearrangements should be more common than interchromosomal

rearrangements; 2) predicted translocations should be kept to a minimum, since these should be less common than segmental deletions or inversions; and 3) each diploid progenitor genome segment should have no more than 2 matching duplicated segments in its paleotetraploid counterpart. Although genome rearrangements are known to be more frequent following WGD events, the subgenome assignment representing the fewest predicted rearrangements is more desirable when inferring evolutionary history.

The algorithm developed to assign these subgenomes worked as follows (Fig 2.1): first, all syntenic blocks from the paleotetraploid species identified using MCScanX were placed into an unordered list. Then, starting with the first gene on the "reference" (diploid, non-paleotetraploid) chromosome, all 'tetraploid' blocks with genes that are collinear with that reference gene are added to the putative list of subgenomes. If 2 blocks match the first gene, each block is assigned to a separate subgenome. Then, for each subgenome assignment, all tetraploid blocks that share a chromosome with those first-assigned blocks are added to the same subgenome assignment as those first-assigned blocks (i.e. assuming the fewest translocations). Next, going down the reference chromosome, each gene on the reference chromosome is checked to see if any unassigned blocks match that reference gene. If so, that reference gene is checked for any other blocks match that gene already on the opposite subgenome, or if another block fits that spot with fewer conflicts (i.e. fewer blocks whose assignment to that subgenome would produce a situation where blocks from the same chromosome are assigned to opposite subgenomes). That block is then assigned to the appropriate subgenome, and the rest of the chromosome is assigned in this way.

This resulted in assemblies of 2 subgenomes for maize to the 10 sorghum chromosomes, and 2 subgenomes for soybean to the 11 common bean chromosomes (Fig 3.1). Importantly, the

assignment to ‘subgenome 1’ or ‘subgenome 2’ is arbitrary and does not imply, for instance, that maize subgenome 1 from sorghum chromosome 1 came from the same ancestral subgenome as maize subgenome 1 from sorghum chromosome 2. While biases like fractionation, expression, TE insertion, or methylation bias can be determined without phasing blocks into chromosome-scale reconstructions in this way (Zhao et al., 2017), and the basic MCScanX output does show multiple syntenic alignments in an unphased manner which allows for these kinds of analyses, the phasing of syntenic blocks into parsimoniously-determined ancient pseudochromosomes allows for the determination of long-distance, persistent, chromosome-scale trends across these subgenomes. For most subgenomes in maize, only 1 or 2 maize chromosomes were needed to cover the entire sorghum chromosome. For soybean, however, often several soybean chromosomes comprised each subgenome matching a *Phaseolus* block. For example, common bean chromosome 3 had 42 and 50 different breakpoints where two different soybean chromosomes matched the reference chromosome at adjacent genes. In contrast, the highest number of breakpoints in maize was 14 and 20 for chromosome 8 in sorghum (Table 3.1). The maize subgenomes covered less of the imputed ancestral chromosome, however, while the soybean subgenomes covered more of their respective reference chromosomes. Two major exceptions to this observation were found: one at one end of *Phaseolus* chromosome 2, where most of the chromosome end had no match in one of the soybean imputed subgenomes; and in the middle of *Phaseolus* chromosome 8, where a large portion of one of the soybean subgenomes had no match. These may indicate large deletions in segments of the soybean genome, or insertions in the *Phaseolus* genome, after the emergence of their common ancestor. Although the syntenic blocks identified in the first step of this research had many gaps, over 95% of all

Phaseolus reference genes were covered by either a gene or a gap in a block, while anywhere from 38.9% to 61.5% of genes of sorghum genes were covered (Table 3.2).

Comparing deletions of genes in reconstructed subgenomes

The primary measure of whether there is a bias or 'dominance' between subgenomes in a paleotetraploid is frequency of deletions between the subgenomes (Freeling et al., 2012). Since the subgenomes for soybean and maize were reconstructed on a chromosome-scale level, the frequency of gene deletion for each segment between a pair of subgenomes can be directly compared. To accomplish this, genes from the reference chromosome were removed from the list of subgenome assignments (i.e. those genes which were not counted as 'covered by blocks' in Table 3.2) and sliding windows of 100 genes were compared between subgenomes. The number of genes in the reference that did not have a match in the soybean or maize subgenome per 100-gene window was counted (Fig 3.2). These deletion levels per 100 genes were generally more divergent in maize than in soybean. Wilcoxon rank-sum tests comparing the overall level of deletions between subgenomes showed that all *Z. mays* subgenomes showed highly significantly different levels of deletions per 100-gene sliding window across entire reconstructed chromosomes, with between 5 to 11 genes deleted in the more-fractionated genome per 100 genes, on average. In *G. max*, by contrast, although many chromosome-level comparisons between the subgenomes were significant, deletion differences per 100-gene window ranged from 0 to 3 at the most, with chromosomes 3, 5, 6, 10, and 11 showing non-significant differences (Table 3.3).

Comparing expression in young leaf tissue between subgenomes in maize and soybean

Another common characteristic of Class I paleopolyploid genomes is a higher level of expression or broader expression of genes from one of the subgenomes. This is purported to lead to a relaxation of selection on the lesser-expressed gene copy, and hypothetically leads to the preferential loss of one of the two (or more) copies of a duplicated gene (Emery et al., 2018; Innes et al., 2008; Panchy et al., 2016; Semon and Wolfe, 2007). Thus, while deletions of duplicated genes and genome segments have occurred in maize and soybean, detecting ongoing fractionation and bias therein requires examining expression levels of these genes. Young leaf tissue RNAseq data was obtained for both of these species at similar growth stages: *G. max* cv. Williams 82 expression data for all tissues was downloaded from NCBI SRA run SRR03738, and NCBI SRA experiment SRR5368994 was downloaded as a representative B73 expression dataset in young leaf tissue. The RNAseq data were mapped using STAR (Dobin et al., 2013) to the v3 maize annotation and the v2 soybean annotation. For each gene model, the TPM (transcripts per million) was calculated, and an average of the $\log_2(1+TPM)$ (Friedman et al., 2006) for 100 genes in a sliding window across each subgenome for each species was plotted (Fig 3.3). In this case, expression between subgenomes tracked quite closely in *G. max,* but was relatively divergent in *Z. mays* subgenomes. A Wilcoxon rank-sum test comparing expression differences between subgenomes in 100-gene sliding windows showed that the sum of $\log_2(1+TPM)$ gene expression in *Z. mays* differed by between 0.404 to 11.915, and between 0.005 and 0.265 in *G. max*. Thus, while bias in deletions was clear and present in *Z. mays* across all reconstructed ancient chromosomes but not in *G. max*, leaf expression levels showed no bias in *G. max* and a slight but inconsistent bias in *Z. mays*.

Examining bias in cytosine methylation state between ancient subgenomes

While pre-existing biases in subgenome expression levels can remain in contemporary paleotetraploid genomes or result in biased deletion of gene copies, other non-sequence characteristics of genomes have been hypothesized to also contribute to bias in fractionation between ancient subgenomes. One example is cytosine methylation, which, while often associated with non-genic sequence content and heterochromatin, can potentially result in changes in expression or lead to gene silencing (Gehring and Henikoff, 2007). As such, investigating whether methylation levels differ in a broad sense between the subgenomes identified here may elucidate whether methylation appears to be associated with other signatures of subgenome bias or dominance. Cytosine methylation is often variable within and among sequences and samples (Niederhuth et al., 2016; Song et al., 2005), but genes can be classified into different categories of gene body methylation based on the general patterns and relative levels of CG methylation of the three different methylcytosine sequence contexts: CG, CHG, and CHH (Bewick and Schmitz, 2017; Niederhuth et al., 2016). Classifications of *Z. mays* and *G. max* genes were obtained from (Niederhuth et al., 2016) which binned all genes into one of four categories based on the methylation patterns in their coding regions: Unmethylated (UM), CG methylated (CG), CHG methylated (CHG), and CHH methylated (CHH). Next, pairs of genes from opposite subgenomes (i.e. genes that have been retained in 2 copies) in soybean and maize were compared to see if their methylation states were similar or different (Figs 3.4 and 3.5). Soybean showed a distinct preponderance of UM-UM gene pairs (Fig 3.4), leading to a marked increase in the percent of gene pairs in soybean with the same methylation state when compared to maize (Fig. 3.5). Wilcoxon rank-sum tests for the difference in the amount of unmethylated genes between each pair of subgenomes for each chromosome were performed (Table 3.3) and showed mixed results with 3 chromosomes of *Z. mays* subgenomes displaying more than 10%

difference in unmethylated genes between subgenomes, and 1 pair of *G. max* subgenome chromosomes showed more than a 10% difference. The *Z. mays* subgenomes had more significant differences as shown by lower p-values.

Discussion

An algorithmic, parsimonious approach to identifying ancient subgenomes

With an abundance of publicly available high-quality, chromosome-scale reference genomes available for an ever-increasing number of crop plants (Bolger et al., 2017; Michael and Jackson, 2013), the opportunities for reconstructing the evolutionary history of some of the most economically and historically important plants for human civilization are readily at hand. Even well before reference genomes were available for many crops, the importance of ancient polyploidy in the development of plant genomes was well recognized (Bowers et al., 2003; Comai, 2005; Paterson, 2005; Semon and Wolfe, 2007; Shaked et al., 2001; Van de Peer et al., 2009; Wolfe, 2001). Early studies in maize, soybean, tomato, wheat, millet, and more showed that large contiguous portions of chromosomes were not only shared between species, but duplicated collinear segments between chromosomes within a genome could be identified, indicating that repeated ancient polyploidies followed by genome rearrangements are not only ubiquitous but likely contribute to diversification and speciation (Anderson et al., 2006; Bonierbale et al., 1988; Devos, 2005; Han et al., 2011; Lin et al., 2010; Uhl, 1992). Furthermore, it became clear with EST (expressed sequence tags) and cytogenetic data that genome rearrangements and deletions of duplicated genes and genome segments often followed polyploidy, reducing these genomes back to a diploid inheritance mode – often denoted diploidization (the process of reduction to diploid inheritance/pairing) and fractionation (the process by which duplicated genes/segments are lost) (Gill et al., 2009; Innes et al., 2008; Lin et al., 2010; Lou et al., 2012; Parisod et al., 2012; Soltis and Soltis, 2012). This greatly complicates the identification of duplicated segments and is a significant obstacle to understanding and visualizing ancient segmental and whole-genome duplications. What is needed, then, is a

consistent method for reconstructing the ancestral, pre-duplication state for duplicated genomes. Typically, this has been accomplished in maize or soybean by aligning orthologs via a basic synteny dot-plot and manually assigning genome segments to one or another putative subgenome arising from a paleotetraploidy (as there are 2 subgenomes created by a tetraploidy event) (Schnable et al., 2011; Wang et al., 2017a). While this manual dotplot-based method can give usable results, it is severely hindered or entirely obviated by genomes which may have undergone numerous rearrangements post-diploidization or which have extremely similar characteristics.

Thus, this study partially aimed to present a consistent, deterministic, and algorithmic method for using contemporary high-quality genome annotations to model preduplication genomes and allow for downstream analyses of the ancient subgenomes arising from these duplications. The method only requires chromosome-level assembly and annotations of gene models from a species with a putative paleotetraploidy, full annotations and assembly for a closely-related species lacking that paleotetraploidy, and a few common bioinformatic tools like BLAST and MCScanX. It assumes broad synteny between the related species, and that the fewest possible translocations, breaks, fusions, or breakages of chromosomes occurred while still allowing for these events. The algorithm uses these precepts to build a block-by-block tiled reconstruction of two ancient subgenomes using the non-duplicated related species' chromosomes as a reference. Fig 3.1 shows the imputed subgenomes for maize and soybean and gives clues to some key differences in the makeup of these subgenomes for these two very divergent species. One notable shortcoming of this method is a lack of accuracy metrics, as this is not, for instance, a maximum-likelihood method. Thus, there is no way to compare the subgenomes built with this method to others without experimental or other data in conjunction.

A reproducible version of the algorithm described here, as well as the formatted data for maize and sorghum used in this study, is available for download as a new software package, "TetrAssign", at https://briannadon.github.io/TetrAssign/.

Z. mays subgenomes had fewer rearrangements than G. max

A cursory evaluation of the reconstructed subgenomes for maize and soybean shows that *Z. mays* subgenomes underwent far fewer translocations, breakages, or fusions than *G. max* subgenomes. While the *Z. mays* subgenomes, as illustrated in Fig. 3.1, are generally comprised of one color (one maize chromosome makes up most of the subgenome), the soybean subgenomes are made up of a mosaic of various chromosomes, with dozens of breakpoints within each subgenome chromosome, where the soybean subgenome switches membership from one soybean chromosome to another. Notably, 12 subgenome chromosomes for maize were made entirely of a single maize chromosome, where no such single chromosome exists for soybean. The soybean subgenomes assigned to common bean chromosome 3, by contrast, had 42 and 50 breakpoints, indicating extensive rearrangements. Often, in these breakpoint-heavy regions in soybean subgenomes, the homoeologous segments consisted of just a few genes (a minimum of 5, from the underlying default MCScanX synteny scan settings) from one soybean chromosome before switching back to another soybean chromosome.

There are numerous explanations for why soybean's subgenomes have apparently undergone extensive rearrangement while maize's have remained seemingly intact since they diverged from common bean and sorghum, respectively. Furthermore, it is unclear whether the rearrangements between soybean and common bean's chromosomes were in the soybean lineage or in the common bean lineage. Some cytogenetic evidence suggests that rearrangements were

relatively extensive between common bean and cowpea (Vasconcelos et al., 2015), which suggests that common bean may have experienced numerous translocations, inversions, fusions, and fissions after its divergence from soybean. In this case, common bean would have undergone rearrangement, scattering and scrambling the matching soybean subgenome chromosomes as shown in Fig 3.1 and Table 3.2, even though soybean's genome itself may have had comparatively fewer of these chromosomal rearrangements.

Syntenic alignment of soybean to *Medicago truncatula*, another related legume species, shows considerable rearrangement of soybean's chromosomes in relation to *Medicago*, however, suggesting that indeed soybean has undergone relatively extensive lineage-specific reorganization, consistent with what is generally observed after diploidization (Wang et al., 2017a). It is possible, however, that instead *Medicago* has also undergone significant rearrangements while soybean would be an exception in that it maintained its chromosome arrangement unlike its *Faboideae* relatives. Perplexingly, wild annual soybean, *Glycine soja,* closely related to *G. max,* is generally karyotypically indistinguishable from other wild perennial *Glycine* species, indicating little rearrangement or dysploidy within the *Glycine* genus, with the exception of a few tetraploid or aneuploid cytotypes like in *Glycine tabacina* (Singh et al., 2001). Even if soybean had undergone these chromosomal changes and not common bean, however, the question remains as to why these legume species have undergone considerable chromosomal reorganization while maize and sorghum did not. Chromosomal rearrangements are thought to be a major source of speciation events and are ubiquitous throughout the evolutionary history of plants (De Storme and Mason, 2014; Uhl, 1992). These often result in abnormalities in pairing and segregation of chromosomes in meiosis, making cytotypes with translocations, fusions, or deletions sometimes incompatible and sexually isolated from other members of its species

(RenÉE Orellana et al., 2007; Winterfeld et al., 2018). Furthermore, plants in unfavorable environments or invasive cytotypes often show variations in their chromosome count and arrangement leading to adaptation to these environments, and these invasive or 'colonizing' plants tend to be selfers – which soybean and common bean are and sorghum and maize are not (Cheptou, 2012; Hiesey, 1966). Thus, some evolutionary or selective pressure may have led to relatively extreme rearrangement of ancestral chromosomes in soybean and little such rearrangement in maize, implying highly divergent evolutionary histories for these crop plants.

Evaluating bias between subgenomes in maize and soybean

As noted previously, the assignment of subgenomes to 'subgenome 1' or 'subgenome 2' in maize and soybean were arbitrary per reference chromosome, which means that a predicted subgenome chromosome belonging to 'subgenome 1' is not necessarily descended from the same ancestral subgenome as another chromosome assigned to 'subgenome 1'. Other studies have, in the past, assigned different subgenome chromosomes to one or another consistent subgenome via measuring biased fractionation (brassica, maize). In these species, which could be considered Class I paleopolyploid species, differing subgenomes have often been defined by determining which genome segments or chromosomes showed more or less deletion of homoeologous genes. This could be accomplished in maize quite easily, as demonstrated (Fig. 3.2a and Table 3.3), which show that the maize subgenomes were differentiable by their deleted gene content per 100 gene window. Soybean, on the other hand, showed far fewer differences in deleted gene content (Fig 3.2b and Table 3.3), and thus attempting to assign subgenome membership based on biased fractionation (deletion) would have been futile. This indicates that soybean's most recent tetraploidy event may have been a class II event, or more like an autopolyploidy or segmental allopolyploidy than a "true" allopolyploidy, as was maize's most recent WGD.

Interestingly, earlier evidence from cytogenetic studies and the first published draft soybean genome suggested that soybean's 8-13 My old WGD event was an allopolyploidy. For example, a study of centromeric repeats (dubbed CentGm-1 and -2) in soybean before the draft soybean sequence was published inferred that these repeats were not homogenized between putatively homoeologous sites and chromosomes, which suggests allopolyploidy (where homoeologous chromosomes do not pair and thus do not recombine or exchange DNA) (Gill et al., 2009). In an autopolyploid, presumably, recombination is expected to homogenize these sequences to some degree due to extensive tetrasomic pairing between homeologs. However, more recent evidence contradicts this claim. Firstly, centromeric or pericentromeric sequences are known to show little to no recombination in plant species (Belling, 1912; Bhakta et al., 2015; Riley et al., 2009; Vincenten et al., 2015), for which recombination is known to often be limited to just one or two chiasmata per chromosome pair due to higher interference than in e.g. animals (Mather, 1940; Muller, 1916; Riley et al., 2009; Sturtevant, 1915). Secondly, this earlier study suggested via comparison that the observed homogenization of centromeric repeats in maize, despite strong evidence of allopolyploidy therein (Woodhouse et al., 2010), was because of extensive DNA exchanges, conversion, or chromosome rearrangements. However, this research showed that in fact syntenic alignment of maize to its relative sorghum paints a picture of ancient maize subgenomes which have undergone little to no reorganization. Thus, while these two soybean centromeric repeats suggest allopolyploidy, a closer investigation of the entirety of the soybean genome enabled by high quality sequence data instead favors a segmental allopolyploidy or autopolyploidy. It is, however, impossible to know if soybean arose from two subgenomes from the same cytotype, two subgenomes from the same progenitor species with differing cytotypes, or two different but closely related species altogether, as the diploid

progenitor(s) that gave rise to the paleopolyploid *Glycine* genus are likely long extinct (Singh, 2016).

Loss of duplicate genes is not only an outcome of bias between subgenomes in a class I paleopolyploid but is also often used as the defining feature of the disparate subgenomes. While deletion of duplicate genes and genome segments is thought to occur relatively rapidly post-duplication in many cases (e.g. in wheat, brassica, Arabidopsis, and *Mimulus*) pre-existing and maintained differences in genome characteristics can potentially mark genes or genome segments for deletion (Edger et al., 2017; Garsmeur et al., 2014; Mandáková et al., 2010; Shaked et al., 2001; Subramaniam et al., 2013). It has been proposed that differences in expression levels in two homoeologous genes between two subgenomes can eventually lead to 'dominance' in one copy, where the higher-expressed copy takes over the function of both original copies, and the lesser-expressed copy experiences relaxed selection and accumulates mutations leading to its, deletion, neofunctionalization, or subfunctionalization (Freeling et al., 2015; Freeling et al., 2012; Lynch and Force, 2000; Wang et al., 2014). In the former scenario, the gene becomes an unexpressed pseudogene or is deleted entirely, and in the latter two cases, the gene assumes a subset of the original functions of the nonduplicated gene (i.e. a gene with functions A+B+C becomes B+C) or gains a new function altogether. Differences in expression between homoeologous gene copies can thus result in a bias in gene loss and can greatly affect the evolution of a genome. Comparing leaf expression levels between soybean and maize subgenomes showed similar patterns to those of gene deletion levels, with some important differences. While the maize subgenomes showed consistent and easily identifiable bias in deletions per 100 genes, expression levels per 100 genes across a subgenome did not show a similar pattern. Despite larger identifiable expression differences via Wilcoxon rank-sum tests

(Table 3.3), individual differences at specific positions along reconstructed subgenomes showed less of a consistent bias in expression, though many windows along these subgenomes in maize did show bias (Fig 3.S1). Informatively, in all 9 cases where there was a statistically significant chromosome-level difference in leaf expression level in the maize subgenomes (for all chromosomes except Sb04, where there was not a significant expression bias), the more-expressed subgenome also experienced fewer deletions. This was not true of soybean subgenomes, however, where in some cases the more deleted genome was very slightly more expressed, such as in those aligned to common bean chromosome 2 (Table 3.3). Pairs of genes between each soybean and maize subgenome were then considered, where both members were present in both subgenomes (i.e. pairs of genes that were present in 2 copies and were not putatively deleted in one or the other subgenome). These pairs (12,707 total in soybean and 3,201 in maize) were classified as "different" if their expression was >2-fold different in young leaf tissue (see methods) or "similar" if not (if there was no detectable expression, they were classified as "NA"). This showed that maize had many more pairs of genes that were different in their expression than soybean (Fig 3.3).

Differences in expression between gene copies arising from WGDs are hypothesized to presage fractionation or deletion of the less 'dominant' gene copy (Edger et al., 2017; Hollister et al., 2011). Some observational and experimental evidence indicate that expression bias between homeologs is established early on – perhaps within a single generation after hybridization – and that these differences eventually lead to biased fractionation of genomes (Edger et al., 2017; Freeling et al., 2012; Mandáková et al., 2010). However, other non-sequence differences between subgenomes in polyploids may also lead to changes in expression and eventually changes in gene function or gene silencing (Edger et al., 2017). Cytosine methylation is one

such epigenetic mark that can lead to changes in chromatin structure and function, and its importance in determining the evolution and adaptation of genes and genomes is still not fully understood (Chen, 2007; Lukens et al., 2006; Salmon et al., 2005; Scheid et al., 2003). Although much of the evolutionary functions and consequences of cytosine methylation have yet to be determined (Bewick et al., 2016; Bewick and Schmitz, 2017; Zhang et al., 2010), some evidence shows that it can have effects on expression patterns, function, and evolutionary fate of genes (Shaked et al., 2001; Zhang et al., 2010). Furthermore, other evidence shows extensive genome methylation re-patterning follows WGD and diploidization events (Salmon et al., 2005; Song and Chen, 2015; Wang et al., 2014; Xu et al., 2018). Understanding how overall cytosine methylation levels and patterns are different between imputed ancient subgenomes in soybean and maize could therefore give clues as to how the evolutionary trajectories of these two species might differ. Often, simply measuring and comparing cytosine methylation levels across a gene body is not enough to detect any meaningful differences between genes or genotypes, so for this study, each gene was assigned a methylation category (Unmethylated/UM, CG gene body/CG, CHG, CHH) according to (Niederhuth et al., 2016). It is important to note, however, that bisulfite sequencing reads used to generate these methylation state assignments only used uniquely-mapped reads, and thus may be confounded or inaccurate in cases where duplicate genes have extremely similar sequences. Comparing levels of unmethylated genes per 100 gene window along the entire length of the reconstructed chromosomes in each subgenome showed some marked differences in maize and soybean (Table 3.3). In maize, once again, the subgenomes had markedly different levels of unmethylated genes, while soybean's subgenomes showed some differences but considerably less than maize. The patterns of differences in unmethylated genes between the two species' ancient subgenomes were not as distinct as the differences in numbers

of deleted genes or expression level of genes, indicating that although methylation may be involved in patterning of genome bias or dominance, it is either quickly lost post-duplication or is not as impactful in the diploidization or fractionation processes as are expression differences or deletion via intrachromosomal recombination (Bewick and Schmitz, 2017; Chen, 2007; Panchy et al., 2016; Woodhouse et al., 2010; Zhang et al., 2010).

Conclusions

This study presents a novel algorithmic method for reconstructing ancient subgenomes that iterates on previously described techniques and adds a level of usability and reliability previously lacking. Furthermore, using this method to compare ancient subgenomes for two of the most valuable paleotetraploid crops in the world, soybean and maize, shows significant differences in their evolution, where maize's subgenomes show marked and consistent differences while soybean's do not. This is in line with mounting recent evidence that soybean's most recent paleotetraploidy may have been a segmental allopolyploidy or autopolyploidy, and not a strict allopolyploidy as was previously suggested. Future work in this regard might include comparing characteristics of ancient subgenomes like repetitive content, noncoding sequences, and chromatin conformation to determine further what kinds of genomic characteristics lead to the differences seen in genomes like maize (class I polyploids), or the lack thereof in genomes like soybean (class II polyploids). In addition, the algorithmic method presented here could be extended and generalized using e.g. recursive methods to allow for ancient hexaploidies, octaploidies, or more to be reconstructed given a suitable representative genome and non-duplicated relative. Furthermore, the integration of more accurate synteny comparisons for one-to-one genome analyses used in this study could be improved by integrating newer algorithms

like QUOTA-ALIGN, or by refining the method via reconstruction of other paleotetraploid genomes.

Table 3.1. Numbers of breakpoints or potential rearrangements within each identified subgenome in soybean and maize. A breakpoint is defined as a point in the 'reference' (sorghum or common bean) chromosome where two adjacent genes in soybean or maize were from two different chromosomes. Subgenome assignments between different reference chromosomes are arbitrary; for example, the *P. vulgaris* chromosome 1's "*G. max* subgenome 1" does not necessarily share ancestry or relation to *P. vulgaris* chromosome 2's "*G. max* subgenome 1".

	P. Vulgaris		*S. bicolor*	
Reference Chromosome	*G. Max* **Subgenome 1**	*G. Max* **Subgenome 2**	*Z. Mays* **Subgenome 1**	*Z. Mays* **Subgenome 2**
1	9	5	0	1
2	8	15	0	0
3	42	50	0	0
4	10	5	0	0
5	5	5	0	0
6	16	13	0	0
7	4	15	2	0
8	5	24	14	20
9	20	2	6	2
10	2	5	1	2
11	8	2		

Table 3.2. Proportions of genes in 'reference' chromosomes (sorghum or common bean) with orthologs matching at least one gene in a soybean or maize subgenome.

Reference chromosome	Total genes aligned	Genes covered by blocks	Percentage genes covered by blocks
P. vulgaris 1	2694	2669	99.07%
P. vulgaris 2	3338	3295	98.71%
P. vulgaris 3	2973	2873	96.64%
P. vulgaris 4	1789	1735	96.98%
P. vulgaris 5	1863	1845	99.03%
P. vulgaris 6	2221	2188	98.51%
P. vulgaris 7	2812	2753	97.90%
P. vulgaris 8	2932	2827	96.42%
P. vulgaris 9	2633	2589	98.33%
P. vulgaris 10	1659	1621	97.71%
P. vulgaris 11	2168	2113	97.46%
S. bicolor 1	5730	3522	61.47%
S. bicolor 2	4416	2332	52.81%
S. bicolor 3	4676	2781	59.47%
S. bicolor 4	3784	2234	59.04%
S. bicolor 5	2436	697	28.61%
S. bicolor 6	2953	1681	56.93%
S. bicolor 7	2373	1148	48.38%
S. bicolor 8	2026	787	38.85%
S. bicolor 9	2687	1418	52.77%
S. bicolor 10	2946	1440	48.88%

Table 3.3. Results of Wilcoxon rank-sum tests for differences in the overall levels of gene deletions per 100 genes, gene expression per 100-gene window, and unmethylated genes per 100 genes between soybean or maize subgenomes. Expression is represented as $\log_2$(1+TPM) (transcripts per million).

Chromosome	Estimated deletion difference	P-value	Estimated expression difference	P-value	Estimated unmethylated difference	P-value
S. bicolor 1	-10.99999	2.20E-16	4.833	2.20E-16	12.00	0.00E+00
S. bicolor 2	9.00004	2.20E-16	-6.149	2.20E-16	-4.00	3.00E-195
S. bicolor 3	-8.99999	2.20E-16	-0.404	1.38E-01	4.00	1.54E-72
S. bicolor 4	-8.00038	2.20E-16	7.790	2.20E-16	9.00	0.00E+00
S. bicolor 5	-6.00005	2.20E-16	1.143	9.56E-04	4.00	8.72E-186
S. bicolor 6	8.99996	2.20E-16	-7.050	2.20E-16	12.00	0.00E+00
S. bicolor 7	-6.00002	2.20E-16	5.064	2.20E-16	8.00	0.00E+00
S. bicolor 8	-4.99995	2.20E-16	11.915	2.20E-16	4.00	4.37E-126
S. bicolor 9	-8.99998	2.20E-16	6.501	2.20E-16	11.00	0.00E+00
S. bicolor 10	-8.00002	2.20E-16	7.290	2.20E-16	9.00	0.00E+00
P. vulgaris 1	-0.99991	1.29E-06	0.063	2.20E-16	1.00	2.76E-04
P. vulgaris 2	3.00003	2.20E-16	0.265	1.33E-03	10.00	1.04E-236
P. vulgaris 3	0.99996	1.85E-06	0.018	1.21E-02	2.00	4.90E-08
P. vulgaris 4	2.00003	2.20E-16	0.056	1.47E-07	-2.00	3.45E-07
P. vulgaris 5	0.00004	6.57E-01	0.023	2.46E-02	1.00	5.87E-02
P. vulgaris 6	0.00006	3.15E-02	-0.065	1.43E-11	2.00	8.69E-07
P. vulgaris 7	-1.00003	2.20E-16	0.024	2.98E-03	4.00	7.23E-44
P. vulgaris 8	-1.99994	2.20E-16	0.036	1.30E-04	4.00	1.80E-42
P. vulgaris 9	-1.00002	2.72E-11	0.062	2.18E-14	3.00	1.24E-44
P. vulgaris 10	-0.99997	9.52E-03	-0.093	5.07E-12	4.00	1.57E-08
P. vulgaris 11	0.00004	2.40E-01	-0.005	7.07E-01	4.00	1.59E-16

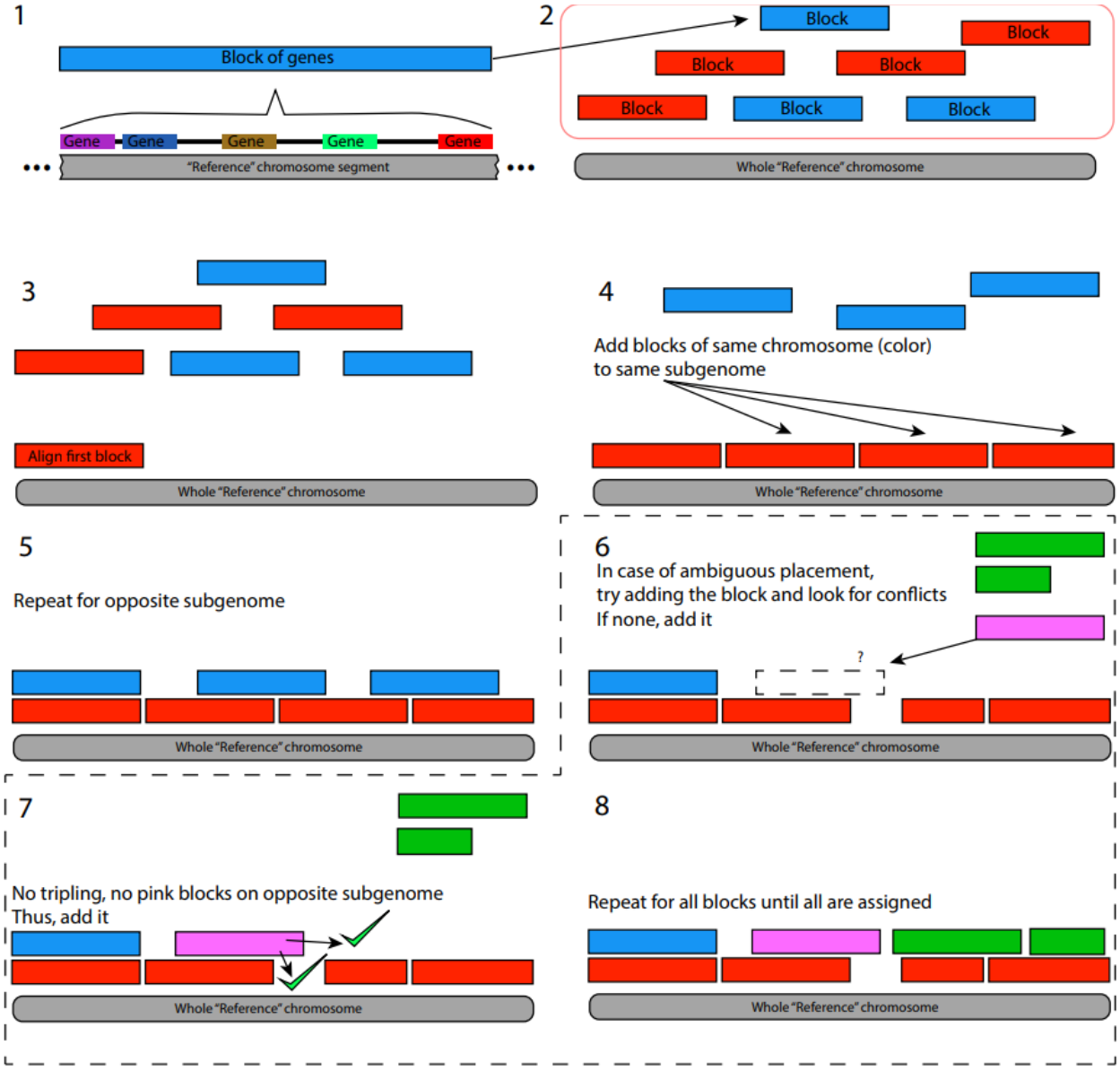

Figure 2.1. Flowchart of subgenome assignment algorithm ("TetrAssign"). 1) All syntenic block alignments from the "tetraploid" species (e.g. soybean) are collected in step 2). 3) The first block matching the first gene on the reference chromosome is added arbitrarily. 4) all blocks matching that chromosome (e.g. soybean Chr01) are added to the same subgenome. 5) This process is repeated for the opposite subgenome. 6) If a new chromosome is encountered, and/or there is no matching gene on the opposite subgenome (see how the pink block would overlap a segment where there is no assigned block for the bottom subgenome), try adding the block. 7) if

there are no conflicts, e.g. no assignments where three blocks would be assigned to one position or this assignment would not place a pink block on the opposite subgenome of another pink block, add that block in that position. 8) repeat steps 6-8 if needed for all blocks until exhausted.

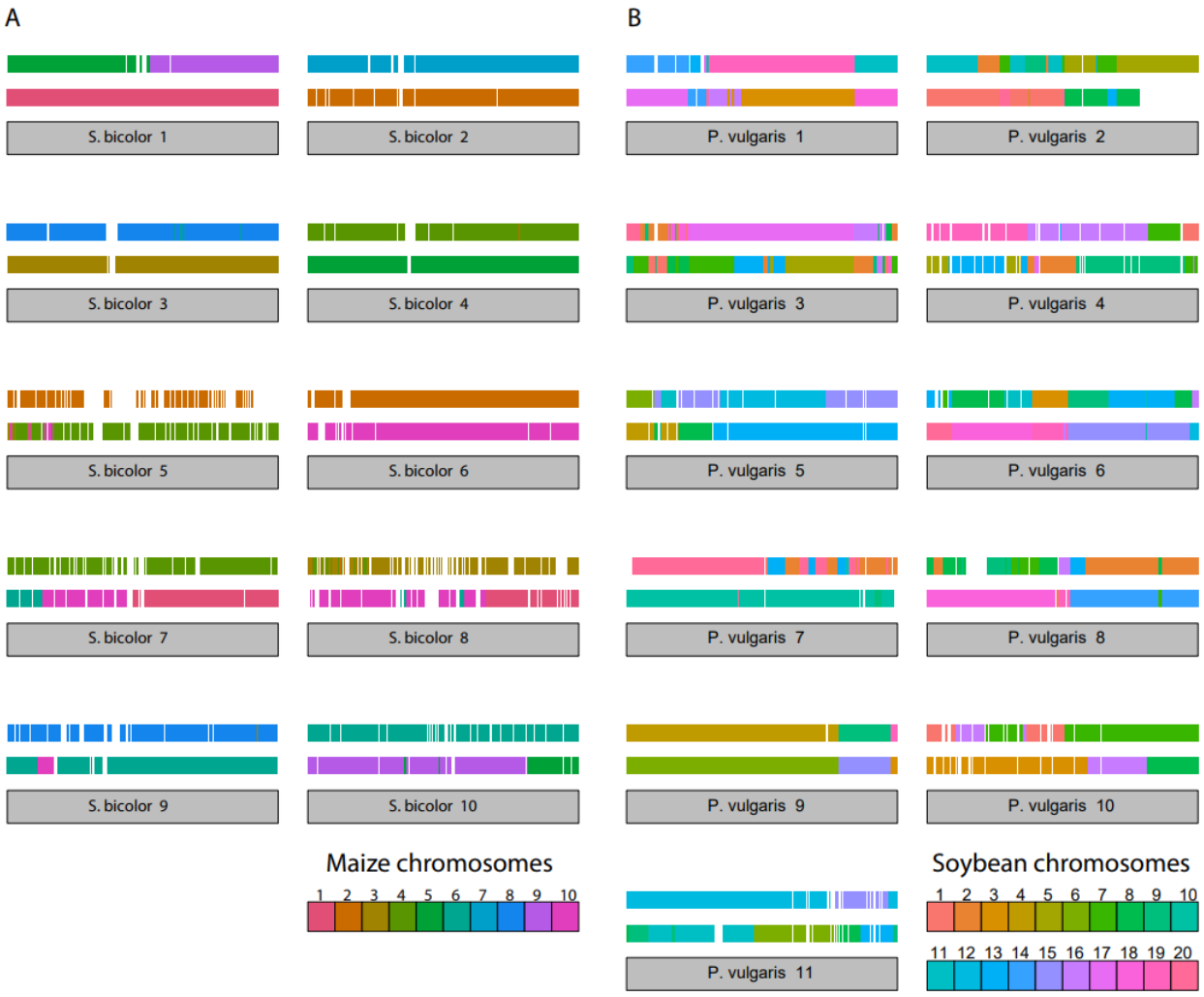

Figure 3.1. Representations of subgenome assignments in maize (A) and soybean (B). Colors represent different maize or soybean chromosome segments assigned to each subgenome. Color assignments are shown in the bottom right corner of each panel, and chromosomes were assigned left to right. Subgenome "1" is on the bottom of each pair and "2" is on the top.

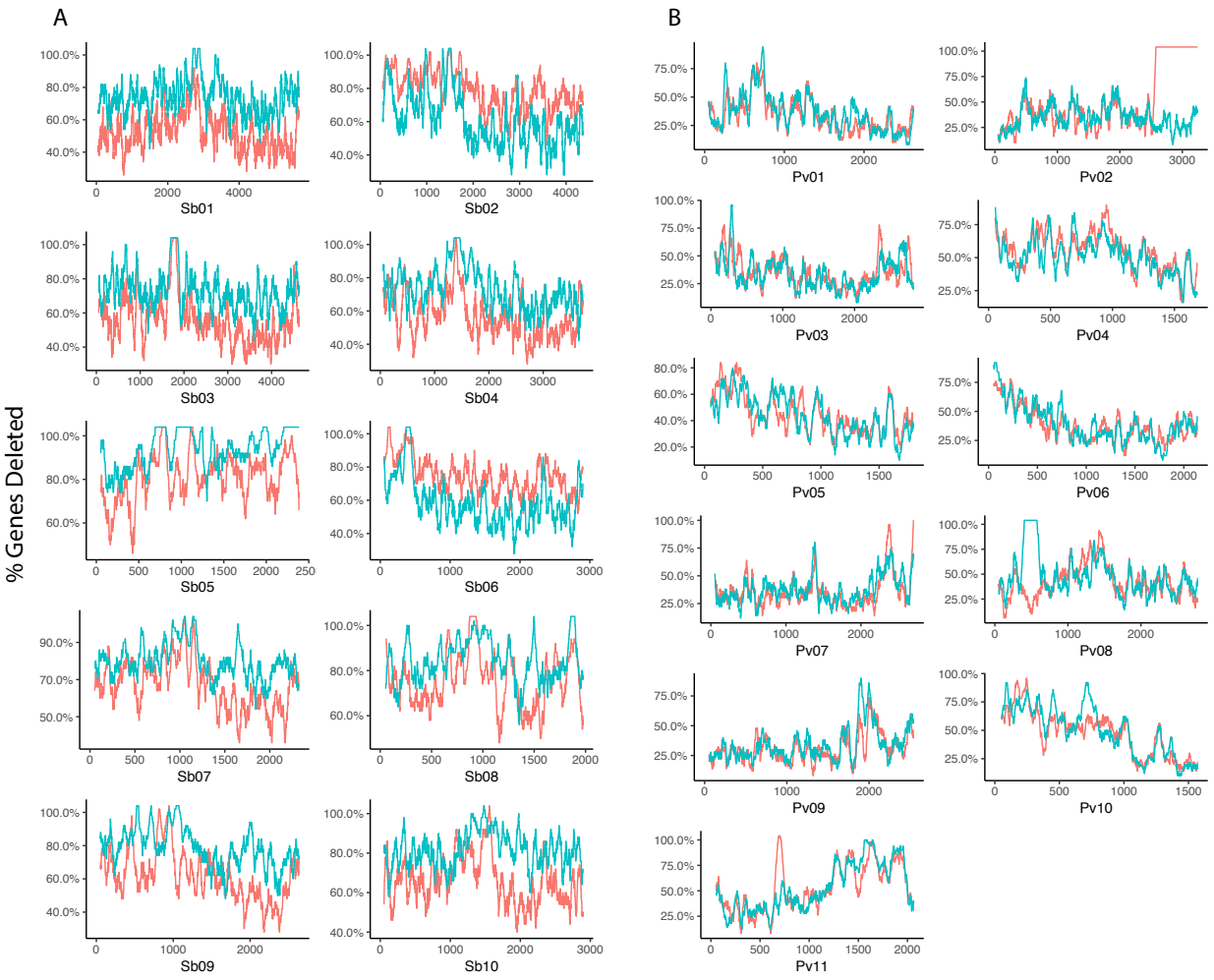

Figure 3.2. Percentages of genes from each (A) maize or (B) soybean 'subgenome' with no ortholog to (A) sorghum or (B) common bean, or putatively deleted genes. The percent of genes deleted was calculated in a 100-gene sliding window with a step of 1, and thus the first approximately 50 genes have no value. Red lines represent subgenome 1 and blue lines represent subgenome 2; these designations are arbitrary per-chromosome. Sections of genes with 100% deletion indicate putative deletion blocks. Each point on the x-axis represents a gene in either common bean or sorghum.

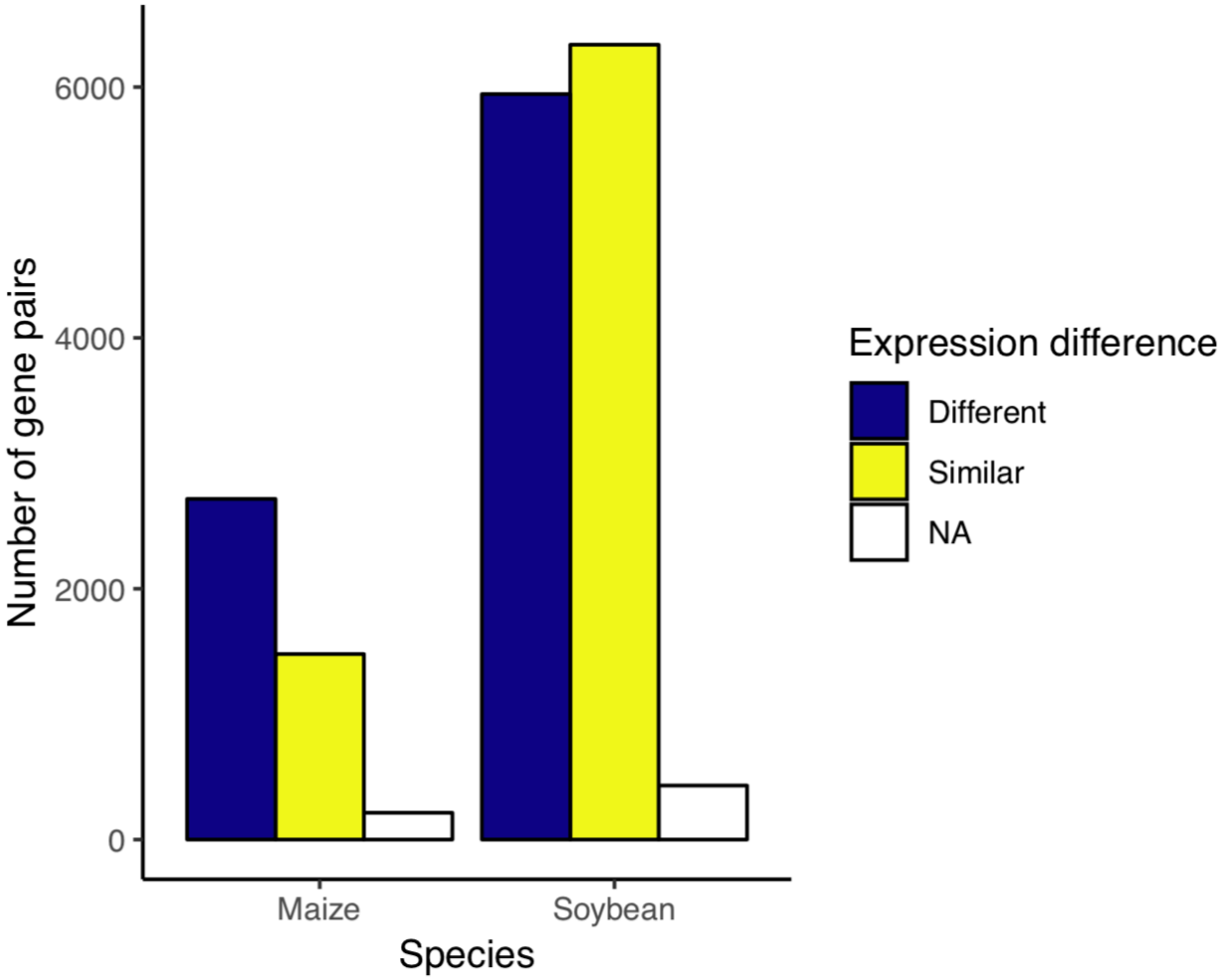

Figure 3.3. Numbers of pairs of genes in maize (3,201 pairs) or soybean (12,707) with significantly different (>2-fold) expression difference in young leaf tissue between gene pairs from opposite subgenome assignments. The "NA" category is comprised of genes with no detectable significant expression in this tissue.

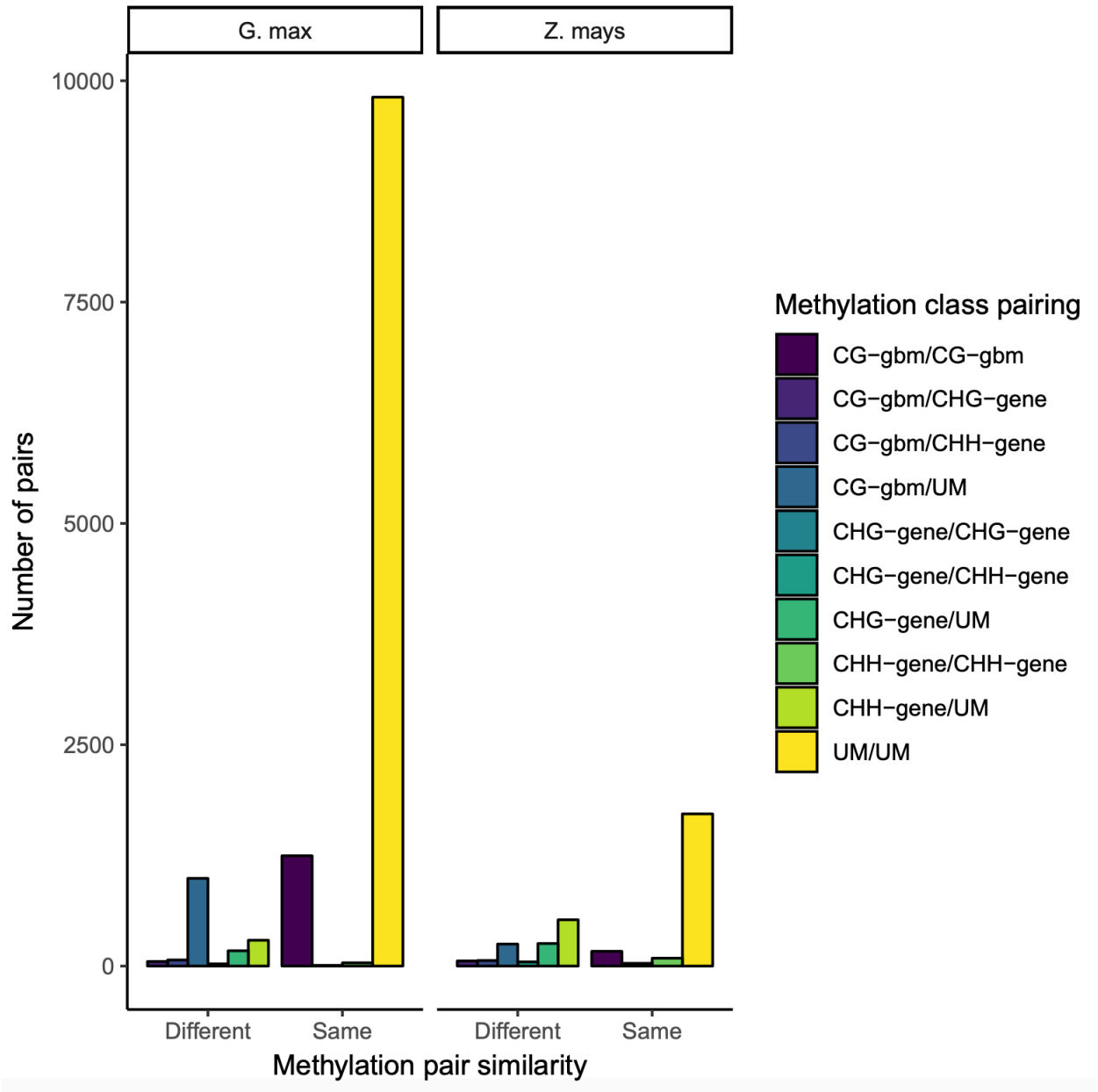

Figure 3.4. Counts of gene pairs with each possible pairing of 4 methylation categories (unmethylated "UM", CG gene body methylated "CG-gbm", CHG methylated "CHG-gene", and CHH methylated "CHH-gene") between each subgenome for soybean (*G. max*) or maize (*Z. mays*). There were 12,707 pairs of soybean genes and 3,201 pairs of maize genes in total.

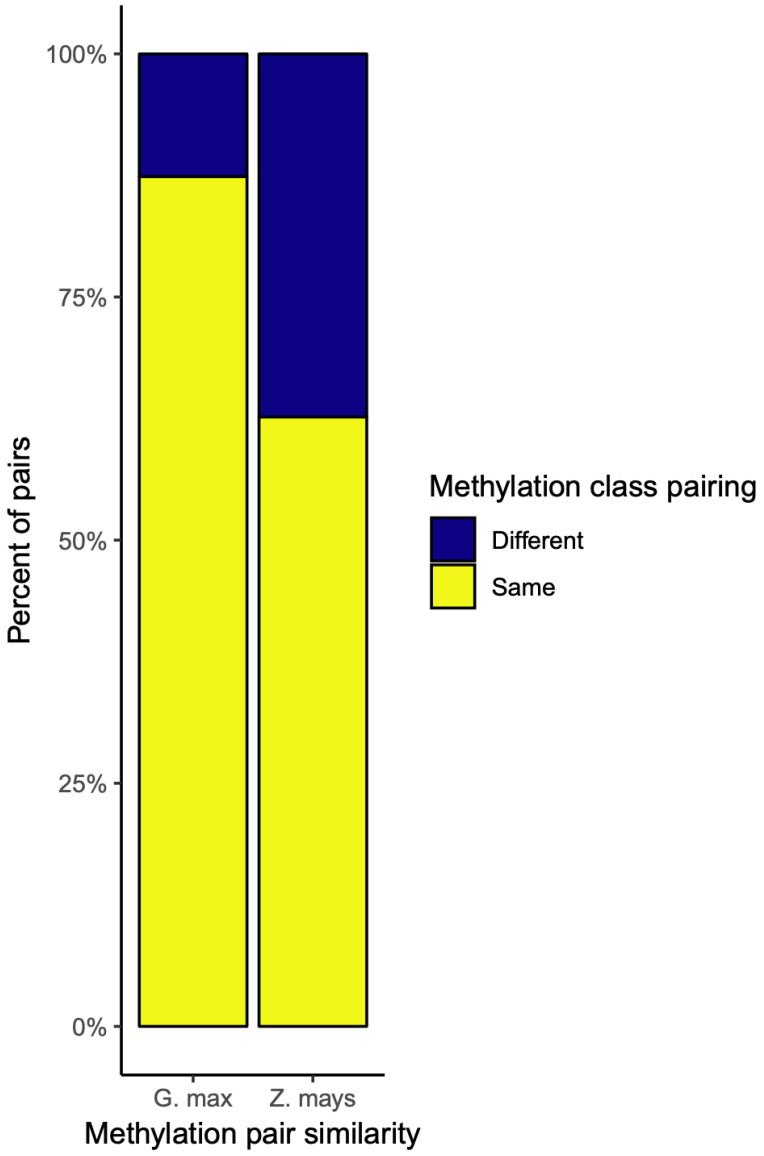

Figure 3.5. Percent of pairs of genes from opposite subgenomes within soybean (*G. max*) or maize (*Z. mays*) with similar (e.g. both are CG-gbm) or different (e.g. one is UM and the opposite is CG-gbm) methylation states. There were 12,707 pairs of soybean genes and 3,201 pairs of maize genes considered here.

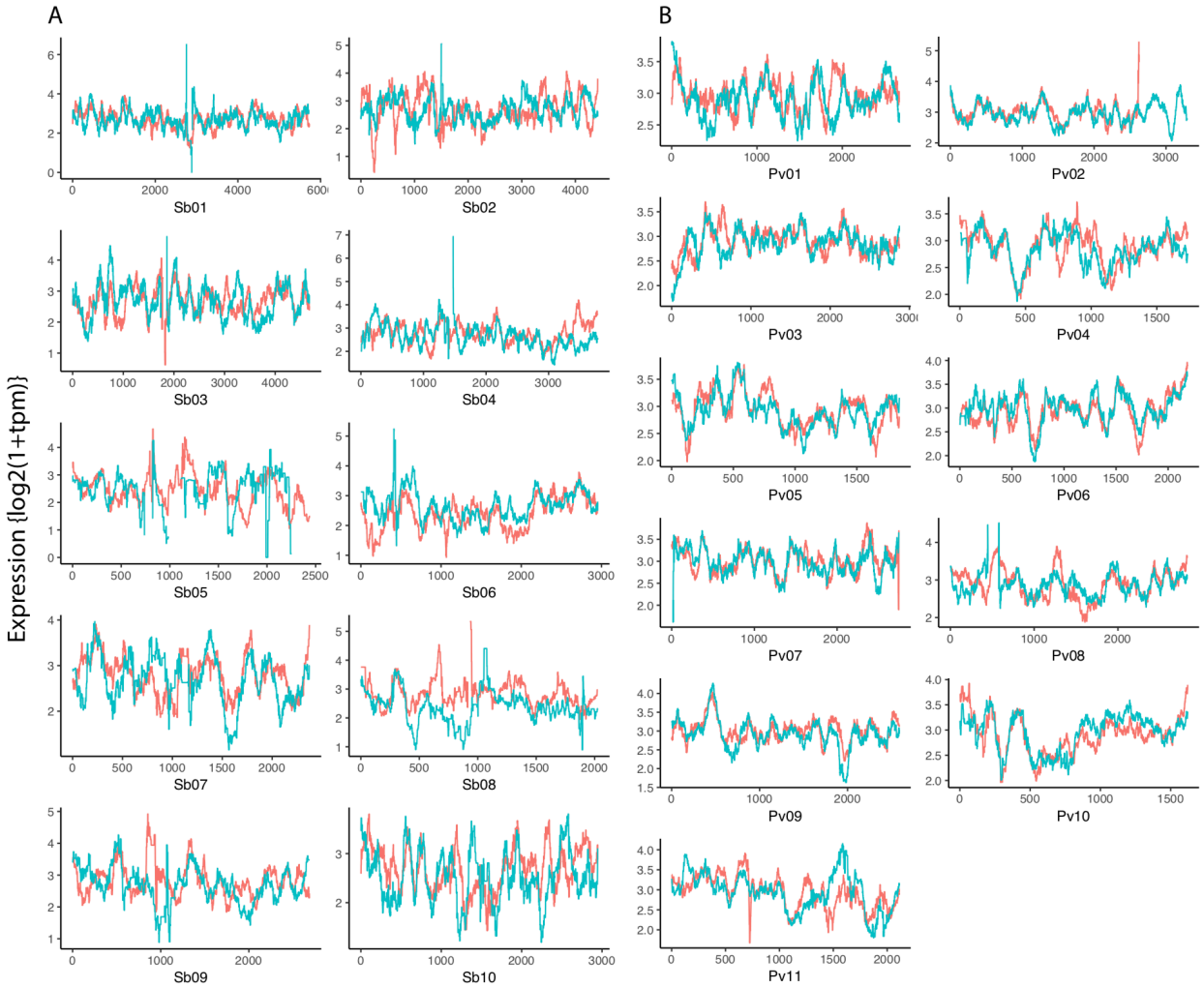

Fig 3.S1. Expression of genes assigned to each subgenome in A) maize or B) soybean across reconstructed chromosomes in young leaf tissue. Expression was measured with $(1+\log_2(\text{TPM}))$.

CHAPTER 4

A HISTORY OF GENE DUPLICATION AND DELETION IN LEGUMES REVEALED BY PHYLOGENETIC ANALYSIS OF GENE FAMILIES

Introduction

Legumes (*Fabaceae*) are one of the most diverse and economically important clades of flowering plants on earth. Among them are included many of the most highly valued crop plants that drive economic activity across the globe: soybean, alfalfa, peanuts, peas, common beans, clover, lentils, carob, mesquite, and more (Graham and Vance, 2003). Notably, legumes are highly valued for their especially high nutritional protein content, owing to their ability to form symbiotic relationships with soil bacteria (rhizobia), which fixes nitrogen in the roots in a nodule (De Faria et al., 1989; Polak et al., 2015). The contribution of legumes to food production around the world is difficult to overstate, and with the prospect of a world with a human population of 9.1 billion or more by 2050, these crop plants will be more important than ever to help meet the nutritional demands of people everywhere (FAO, 2013; Population Reference Bureau, 2007). As such, a deep understanding and appreciation of the genetic and genomic evolution and diversification contained within the legume family will be vital to shaping these legume food and feed crops into crops that can feed 9 billion. With the rapidly expanding public availability of high quality, fully annotated genome data for many different plants, a deeper understanding than ever before of the evolution of the genomes of these legume crop plants is possible.

One notable characteristic of plant genomes is their apparent tolerance for whole genome and segmental duplications. Some 25-30% of extant land flowering plants today are polyploid

(or neopolyploids) (Meyers and Levin, 2007; Salse, 2016; TATE et al., 2005), carrying more than 2 copies of their genomic content in a typical sporophytic cell, varying from 4x to 44x or even higher (e.g. black mulberry *Morus nigra*) (Dandia et al., 1990). Because of the widespread phenomenon of polyploidy across geography and time in plants, a history of ancient polyploidy is apparent in all characterized flowering plant genomes on earth (Adams and Wendel, 2005; Cui et al., 2006; Murat et al., 2017). All flowering plants (angiosperms) share at least two ancient whole-genome duplication (WGD) events: one from the common ancestor of seed plants and one from the common ancestor of all angiosperms (Murat et al., 2017; Soltis et al., 2008). As such, any well characterized plant genome displays evidence of a history of polyploidy or duplication. Even diploid plants show this phenomenon due to the processes of diploidization and fractionation. Diploidization is the process by which previously polyploid plants become diploid over time (Mandakova and Lysak, 2018), and fractionation describes the phenomenon where a previously polyploid genome, after diploidization, deletes duplicate gene copies and reorganizes its duplicate genome segments such that the remnants of this polyploidy remain visible in their genomes in the form of duplicate gene copies and large duplicated segments of contiguous genes (syntenic blocks) (Wang et al., 2011; Wang et al., 2012). The frequency and rate at which duplicated genes are lost to these processes or retained in duplicate for long periods of evolutionary time is known to vary within and between gene families, within and between ancient genomes, and across lineages (De Bie et al., 2006; Garsmeur et al., 2014; Hahn et al., 2007).

Duplication of genes and entire genome is thus critical to understanding how plant genomes evolve, and it is worth investigating how this has affected legume crops. The *Faboideae* (previously "*Papilionoideae*" or "Papilionoid") clade, a subfamily of legumes which

includes many of the most valuable crop legumes produced worldwide like peas, peanuts, common bean, and soybean, shares at least two known ancient genome duplications among all its members (both more recent than the seed plant and angiosperm duplications mentioned above) (Fig 1.1). One of these is presumed to have occurred within the common ancestor of the core eurosid clade, and occurred an estimated 100 to 130 million years ago (Mya) (Tang et al., 2008b; Zheng et al., 2013). This event is suspected to have been a hexaploidy, or more like a triplication than duplication. A more recent duplication event, estimated at about 60 Mya, occurred at the base of the *Faboideae* and defines the members of the clade (Doyle and Luckow, 2003; Renny-Byfield and Wendel, 2014). Other, newer polyploidies are scattered throughout the *Faboideae*, as is common among all flowering plants: cultivated peanut (*Arachis hypogaea*), for instance, is a recent (<10kya) allotetraploid (Kochert et al., 1996; Leal-Bertioli et al., 2015); alfalfa (*Medicago sativa)* is often cultivated in an autotetraploid form (Brummer et al., 1999); and the *Glycine* genus containing the most economically valuable legume in the world, soybean (*Glycine max*), shares a now-diploidized duplication event likely concurrent with the genus' appearance around 8 to 13 Mya (Gill et al., 2009; Schmutz et al., 2010b; Wang et al., 2017a).

Whole-genome duplications are not the only kind of duplication plant genomes experience, however. The fate of duplicated genes after whole-genome, segmental, tandem, dispersed, or other duplications is a topic of great interest to geneticists as a growing body of work over more than a century has illuminated the critical role all kinds of duplication play in genome evolution (DeVries, 1915; King et al., 1998; Lutz, 1907; Magadum et al., 2013; Winge, 1917). Myriad hypotheses and theories have been developed to explain or predict the evolutionary trajectories of duplicate copies of genes. In general, duplicate copies of genes can be retained, deleted, or experience a change in function (Lynch and Conery, 2000). In the case

of retention of duplicates, the dosage balance hypothesis suggests that genes that participate in multimeric complexes or networks need to be present in the proper stoichiometric ratios to yield a functional product, and thus duplicates are retained to maintain the appropriate ratio of components (Birchler and Veitia, 2012; Birchler and Veitia, 2014). In the case where duplicates are deleted, one copy of the gene is either subject to relaxed selection and accumulates mutations rendering it nonfunctional/psuedogenized, or the higher dosage of the two (or more) duplicate copies results in a deleterious phenotype and thus negative selection and eventual removal of one copy. In the final case, duplicate genes can take on subsets of the original non-duplicate parent gene's function (subfunctionalization), where e.g. a gene with functions A+B duplicates into two genes with one gene with function A and the other with function B, or the duplicates can take on entirely new functions to due relaxed selection constraints on one copy (Flagel et al., 2008; Lynch and Force, 2000; Panchy et al., 2016; Xu et al., 2015). There is evidence that one or more of these processes can work on a given duplicated gene family at once, and that there is significant interplay between them (Panchy et al., 2016). Close analysis of gene families, including their sizes, functions, and expansion or contraction over evolutionary time can therefore elucidate the processes that are shaping the characteristics of genes in modern genomes.

There are many ways to define and describe the ways that these genomes have duplicated, such as Ks analysis, analysis of syntenic gene blocks, or best-reciprocal-BLASTing to find orthologous or paralogous genes (Cannon et al., 2006; Wang et al., 2017a). In all cases, high quality genome annotations and complete protein sequences allow for more accuracy and a higher-resolution approach. It is possible with high-quality reference genome annotations to attempt to assemble a comprehensive set of gene families that describe all identifiable paralogs

and orthologs for a given gene within a selected set of species. Analyzing these gene families using phylogenetic or maximum likelihood approaches can allow for an accurate estimation of the evolutionary history of the genomes under consideration.

In this study, the genomes of the economically important species soybean (*Glycine max*), common bean (*Phaseolus vulgaris*), alfalfa relative and model species *Medicago truncatula*, wild diploid peanut (*Arachis duranensis and Arachis ipaensis*), cultivated peanut (*Arachis hypogaea*), and non-legume eurosid out-group species grapevine (*Vitis vinifera*) will be used to construct a set of comprehensive gene families for the *Faboideae* legumes. Each gene family can be represented as a phylogenetic gene tree, and a model-based approach can be used to compare these gene trees with different models of gene family evolution that represent different scenarios of retention and deletion, allowing for the estimation of the evolutionary history of these gene duplicates in these crucial crop species and how they shape these genomes today.

Materials and Methods

Complete protein sequence annotations for *Vitis vinifera* (Genoscope.12X version, phytozome), *Arachis duranensis* (peanutbase.org, version aradu1.0), *Arachis ipaensis* (peanutbase.org, version 1.0), *Arachis hypogaea* (peanutbase.org, version 1.0), *Medicago truncatula* (Mt4.0v1, phytozome), *Phaseolus vulgaris* (v2.1, phytozome), and *Glycine max* (Wm82 a2v1, phytozome) were obtained and filtered to ensure only the primary protein isoform was represented for each gene. In the case of *A. hypogaea*, which unlike the other species included is an allotetraploid, the genome files were split into two files: one with only genes belonging to the A genome (whose progenitor is *A. duranensis*) and one with only genes belonging to the B genome (whose progenitor is *A. ipaensis*). In order to group all genes from these genomes into orthogroups representing complete gene families (all orthologs + all

paralogs), Orthofinder 1.0.6 was run using the default settings (Emms and Kelly, 2015). This resulted in a database of orthogroups describing groups of genes from all combinations of the species that are predicted to represent the complete ortholog and paralog set for a given gene, along with gene trees, reconstructed gene trees, and a predicted species tree derived from the orthogroups data for these eight species.

The resulting reconciled gene trees representing the orthogroups were reformatted for analysis with the ete3 toolkit (version 3.1.1) and treeKO (http://treeko.cgenomics.org/). A set of 15 topology-only phylogenetic trees was created that modeled deletions occurring at every possible branch along the expected gene tree (the 'neutral' or 'full retention' tree). The "neutral" or "full retention" tree models the scenario where there have been two major duplication events among legumes (one ~58 Mya and one~8-13 Mya in *Glycine* only) with no gene loss (Fig 4.3O). Deletions were modeled at the base of the legumes (i.e. a deletion before the legume common ancestor arose), directly after the legume duplication (resulting in single copies for each legume except soybean), and before and after the divergence of each species considered here. Using treeKO (Marcet-Houben and Gabaldon, 2011), a duplication-aware algorithm that measures distances between two phylogenetic trees, distances were calculated between each model tree and each orthogroup gene tree. Distances are represented as a value between 0 and 1, where 0 represents two trees with no differences in topology, and 1 represents two trees that have no similarities in topology whatsoever.

For each orthogroup, the distance between the observed gene tree of the orthogroup to each of the 15 model trees was calculated using treeKO. For each comparison, both the strict and speciation distance were calculated. The strict distance represents the distance between two trees calculated by comparing "pruned" gene trees, where each tree only has one member of each

species, to each other, and assigning a distance based on these strict criteria. The speciation distance, in contrast, takes primarily into account whether two duplicated gene trees have the same species topology, with gene duplication events having the same evolutionary history subsequent to the event being assigned a distance of 0 (i.e. a distance of zero indicates there is no difference save for the duplication). Using these distances, the estimated probability of gene loss in each branch of the species tree was calculated using the mean strict or speciation distance between the model tree that represented a loss in that branch and all orthogroups. For some orthogroups, distances could not be calculated due to these orthogroups containing only one species. These were excluded from further analysis.

To determine what gene classifications and functions are associated with membership in orthogroups whose relationships are best explained by each different duplicate gene deletion model, a Gene Ontology (GO) enrichment analysis was performed. First, publicly available gene ontology terms for 7 of the species available were downloaded as follows: GO terms for genes in *G. max, P. vulgaris, M. truncatula,* and *V. vinifera* were all obtained from annotation info downloaded from Phytozome V12. GO terms for *A. duranensis* and *A. ipaensis* were obtained from peanutbase.org data stores (both from their respective version 1.0 annotations). Predicted GO terms for genes in *A. hypogaea* were not publicly available at the time of this analysis. In order to obtain appropriate gene function prediction via GO terms for cultivated peanut *A. hypogaea*, an Interproscan analysis was performed with default settings and the "-goterms" option, which outputs GO terms along with predicted gene functions and classifications from other analyses. The outputs of this interproscan analysis were combined with all other GO term annotations to create a database of GO terms associated with every gene in every genome in these 8 species. In total, 2280 unique GO terms were included in the analyses, although not all of

these GO terms were represented in the dataset. Then, for each orthogroup, the deletion model (from the 15 described above) that showed the smallest distance between the orthogroup's gene tree and the model tree was assigned as that orthogroup's "predicted model," or the predicted deletion model that best explains that orthogroup's topology. The GO terms that described each gene were associated with their respective orthogroups. Thus, a database was created that lists every gene, along with its predicted GO terms (if any), the orthogroup it belongs to, and the deletion model that best explains its orthogroup assignment. For each of the 15 deletion models and each of the 2280 GO terms, a 2x2 frequency table was constructed that gave the counts of genes that are in orthogroups best represented by that model or not (e.g. "full retention" vs. "all other models") and genes that were associated with that GO term and genes not associated with the GO term. Some tables could not be constructed because there were no genes that matched that GO term, and these were excluded from further analysis. For each of these 2x2 tables, Fisher's Exact test was used to determine if each GO term was significantly enriched for genes belonging to a given deletion model. GO term – model combinations for which the p-value obtained from these Fisher's Exact tests was < 0.05 were considered significant. Due to the method of table construction, an odds ratio < 1.0 represents a GO term enriched in that model.

To address the problem of differential duplicate gene evolution in these gene families from a second approach, the gene families were modeled with a WGD-inclusive Markov birth-death model using the R package "WGDgc". Briefly, this method models each gene family as a discrete-time Markov chain, where each time step includes a probability of gaining a single copy in the family (birth rate) or losing a single copy (death rate). It also includes, on branches were a WGD was expected, probabilities for the resultant two duplicate copies of every gene in the ancestral genome instantly losing one copy (P(1)) or retaining both (P(2), or retention). In this

case, this means two WGDs were included in the model: one for the ~60 My old papilionoid shared duplication event and one for the 8-13 My old soybean-specific event. First, a table of the numbers of genes from each species in each orthogroup was obtained from the standard Orthofinder output. Next, because WGDgc could not be feasibly run on all included orthogroup data (required upwards of 2000 Gb of RAM), the table was divided to only include gene families with 50 or fewer or 20 or fewer members. Furthermore, these two subsets were themselves subsetted into 5000 random families or all included families, resulting in four total datasets. For each of the 4 datasets, a nonlinear optimization method was implemented into WGDgc via "nlminb" by editing the source code to replace the portion which originally used the standard R "optim" method, since "optim" again required too many RAM and CPU resources to use feasibly (likely due to the size of the data). This modified WGDgc package was used to estimate the Maximum Likelihood estimated parameters for birth rate, death rate, and P(1) and P(2) for both WGD events.

Results

Constructing gene families for 6 legume species and grapevine

Gene families were defined for the 6 legumes under consideration for this study (*Glycine max, Phaseolus vulgaris, Medicago truncatula, Arachis duranensis, Arachis ipaensis, Arachis hypogaea*) with grapevine (*Vitis vinifera*) as an outgroup. Although cultivated alfalfa is the species *Medicago sativa*, its close relative *M. truncatula* was chosen because a well annotated and assembled genome was publicly available for the latter, owing to its longtime status as a model legume species, whereas the former lacks such resources. Because *A. hypogaea* is an allotetraploid whose closest living progenitor diploids are *A. druranensis* (the A genome contributor) and *A. ipaensis* (the B genome representative), the genes belonging to the A and B

subgenomes were split so that *A. hypogaea*'s two subgenomes were treated as separate species, *A. hypogaea*-A and –B (Fig 4.1). This study does not treat the extremely recent allotetraploidy in peanut (as recent as 10 kya) as an ancient WGD or a paleopolyploidy, as it has yet to diploidize. Complete primary transcript protein sequences were obtained from Phytozome, Peanutbase, and JCVI for each species for a total of 300,094 genes. Using Orthofinder v2.2.7 (Emms and Kelly, 2015), these proteins were grouped into 25,147 "orthogroups" (i.e. gene families representing all putative orthologs and paralogs of a given gene member) (Table 4.1). Most of the genes from these species were assigned to an orthogroups, with a maximum of 93.7% of genes belonging to an orthogroups in *P. vulgaris* and a minimum of 67.7% of genes in *M. truncatula*. *Medicago* also showed the highest amount of species-specific orthogroups (orthogroups containing members of only one species, or only paralogs and no orthologs) with 1228 genes in 79 orthogroups, although many of these may be misannotated TE-related genes. A BLAST search of *Medicago* annotated genes against RepBase 19.06 with e-value 1E-100 produced 1,924 TE-related genes on chromosomes (or 2,234 including scaffolds), indicating that up to 3.8% of *Medicago* annotated genes were possibly mis-annotated TEs, on top of potentially many more mis-annotated due to incomplete assembly or split gene models. The mean orthogroup size was 8 genes and the median 10 genes, and 50% of genes were contained in an orthogroup with 13 members or more, leading to a somewhat bimodal, right-skewed distribution of orthogroup sizes with modes at 1 and 10 (Fig 4.2). 13,897 orthogroups, or 55.7% of orthogroups, contained at least 1 member from each species. 868 families were single copy only, with exactly one gene copy in every species in the orthogroup. An example orthogroup represented as a gene tree is presented in Figure 4.3, showing a gene family with at least one gene member from each species included in this study.

Estimating the likelihood of duplicate gene deletion using a model-based phylogenetic approach

Like gene families, the orthogroups can be represented as a phylogenetic gene tree (Fig 4.3). Overall, 19,118 of the 25,147 orthogroups could be assembled into coherent and dichotomous gene trees and were used for all further phylogenetic analyses. Since one outcome of gene duplication is the eventual degradation, pseudogenization, and loss of one copy of the duplicated gene (Byrnes et al., 2006; Parkin et al., 2014; Schnable et al., 2012; Schnable et al., 2011), a phylogenetic approach utilizing these gene trees can help to elucidate whether different lineages among the 7 species studied here experienced different rates or frequencies of duplicate gene loss.

To accomplish this, first, a neutral model of duplicate gene evolution among these species was generated as a topology-only (no distances) newick tree (Fig 4.4a). In this model (the "full retention" model), there is no gene loss after any of the duplication events known to have occurred in these lineages (the 'neutral' assumption here). Accordingly, a duplication event at the base of the *Faboideae* is present, along with the *Glycine*-specific duplication event, leading to one copy of the grapevine gene, 2 copies for all legumes except soybean, and 4 gene copies in soybean. To simulate deletions of a single gene at each possible branch in the species tree, 14 additional model trees were constructed, each representing a single gene loss at a single branch along the species tree (Fig 4.4b-o). With each of these topology-only trees available as a model tree to compare against each of the observed gene trees, a measurement of distance was needed to describe the disparity between each or all of the 19,118 gene trees and the 15 model trees.

The typical solution to comparing the difference or distance between two phylogenetic trees, the Robinson-Foulds distance algorithm (Robinson and Foulds, 1981), would not work here because both the model and the observed gene trees are expected to have duplicates of any

or all species within the tree. Thus, TreeKO, a duplication-aware distance algorithm based on Robinson-Foulds was chosen for this purpose (Marcet-Houben and Gabaldon, 2011). Briefly, this algorithm works by recursively decomposing a duplicated gene tree into a series of 'pruned' trees containing a maximum of one member from each species and calculating the distances separating these pruned trees from the target tree (in this case one of the 15 models). For each of the 19,118 orthogroups with gene trees, the treeKO strict distance was calculated between the gene tree for that orthogroup and each of the 15 model trees for a total of 286,770 distance comparisons.

Because each model tree represented one of 15 possible single-branch deletion events along the species tree, the average distance of each model representing a deletion to all 19,118 observed gene trees could be mapped onto the species tree estimated by Orthofinder (Fig 4.5). In this case, it was important that any estimations of the likelihood of gene deletion along a branch take into account the length of that branch, because under a Brownian or neutral evolutionary model it is expected that gene deletion or pseudogenization is more likely along a longer branch by simple chance alone (Butler and King, 2004). Thus, the calculated average distance for each model (i.e. each species tree branch) was divided by the appropriate branch length in the estimated species tree. The branch with the estimated highest estimated likelihood of deletion was the base of the *Faboideae*, and the branches with the lowest estimated likelihood of deletion were those belonging to the base of the *Arachis* A and B subgenome related lineages (Table 4.2). To evaluate a different approach to assessing the varying rates of gene deletion among these species, the one model that minimized the treeKO distance to each orthogroup was assigned to that orthogroup as the best model for that orthogroup. The model corresponding to a deletion in the common ancestor of the *Faboideae* was the best representative model for the most

orthogroups, and the models for a deletion at the base of either of the *Arachis* A or B genome lineages were the best fit for no orthogroups except one. Fig 4.6).

Estimates of rates of post-WGD retention and gene deletion and duplication using maximum likelihood Markov models

A straightforward phylogenetic approach as described above can yield rough estimates for the relative likelihood of different scenarios of gene deletion across a lineage. However, there are other approaches to the question of how likely gene deletions were across a lineage, and how they might be different among different clades within that lineage. One oft-applied approach is Markov models and maximum likelihood methods to estimate the background duplication and deletion rates of genes in a lineage. In short, this involves repeatedly estimating parameters of gene duplication (often denoted λ) and deletion (μ) and comparing these parameters to observed gene family data to calculate how likely each collection of parameters is until the optimum parameters are determined (Jiao et al., 2011; Nei et al., 1997; Rabier et al., 2014). This approach, however, often assumes duplication and deletion are constant across a lineage and does not take into account the instantaneous explosion in gene family size associated with a WGD event, which is also often followed by a rapid contraction (deletions). Thus, in order to incorporate the two known WGD events among the legumes included in this study, a combined approach was necessary.

A working approach to using gene family data to calculate background gene duplication and deletion events along with retention post-WGD has been described in yeast and simulated datasets (Rabier et al., 2014; Tiley et al., 2016). This approach is notably sequence-agnostic; it only uses a dataset that describes a set of gene families and the number of genes from each species in that family. Thus, the gene families identified using orthofinder in this study can be

used with this method to simultaneously estimate a general background duplication and deletion rate for the legumes, and to determine rates of gene retention post-WGD for both of the events in that lineage: the basal ~60 My old *Faboideae* duplication and the ~13 My old *Glycine* duplication. After filtering of data to make computation feasible, the WGDgc R package (Rabier et al., 2014) was modified to tolerate the large legume dataset generated here, and to use a more robust optimization method with non-linear minimization (Gay, 1990).

Since the number of gene families included in this study was large at over 25,000, and there were several families with hundreds of members, data filtering and subsampling was performed to both make computation feasible and to test whether larger gene families or larger samples would affect birth, death, and WGD retention estimates. Thus, subsets of gene families with 20 or fewer members or 50 members or fewer were taken, along with an additional dataset for each where 5000 gene families were randomly selected of those subsets. The results of the maximum likelihood estimation for all four datasets are shown in Table 4.3. In order to obtain an estimate of the birth and death rates over time for the legumes, the evolutionary rates and divergence times of the legumes were adapted from (Lavin et al., 2005) to calibrate a molecular clock for the species considered here. The divergence of the MRCA of *V. vinifera* and the legumes chosen for this study was set at 110 Mya, and the 'ape' R package's chronoMPL function (Paradis et al., 2004) was used to obtain a time estimate per branch unit of approximately .0033 per million years. This resulted in an estimated birth rate of approximately .003 to .005 births per gene per million years and a death rate of .011 per gene per million years (Fig 4.6). Retention rates immediately following the *Faboideae* duplication were estimated to be ~ 0.27 and following the *Glycine* event ~ 0.79 (Fig 4.7).

Gene Ontology term enrichment and association with gene family history or size

An analysis of Gene Ontology (GO) terms attached to genes can give clues as to the broad types and functions of genes based on their structure, homology, and experimentally observed behavior (Ashburner et al., 2000). Associating GO terms with membership in orthogroups of different size, evolutionary history, and topology will help to elucidate how evolutionary history in legumes is tied to gene function. GO term assignments for genes in the species included in this study were mostly publicly available and GO term assignments for *A. ipaensis* and *A. duranensis* downloaded from Peanutbase.org, assignments for *P. vulgaris, M. truncatula,* and *G. max* obtained from Phytozome v13, and GO terms for *V. vinifera* obtained from Genoscope (v12). GO terms for genes in *A. hypogaea* were not available and thus were built *de novo*. The peanut annotations and gene models were analyzed with InterProScan 5 (Jones et al., 2014) using default parameters and GO term output enabled. The resulting GO terms were extracted and compiled such that all GO terms associated with each gene were taken together. This represents the first protein domain-based analysis of gene function in cultivated tetraploid peanut.

To determine broadly what kinds of genes are associated with larger gene family size and this a higher probability of duplication or retention, a set of one-way ANOVA tests was performed with each of the GO terms identified across all the species. Briefly, every gene was assigned a 1 or 0 for each GO term, and the size of the orthogroup (total number of genes) that the gene was assigned to was used as the independent, continuous variable. This resulted in a total of 2280 ANOVA tests on 300,094 genes per test. A total of 215 GO terms were found to be significant predictors of orthogroup size, with 10 terms positively correlated with orthogroup size and the remaining 205 significant terms being negatively correlated with orthogroup size. The 10 terms showing positive correlation to orthogroup size are presented in Table 4.4**.** Terms

like ‘nucleic acid binding’ and ‘protein dimerization activity’ were notable as highly significant terms with strong predicted positive effects on orthogroup size.

To investigate which GO terms were associated with different evolutionary trajectories, the GO terms were associated with deletion models identified in the previous section. A hypergeometric test was performed for each of the 2,280 GO terms for every orthogroup, and GO terms that were significantly enriched or depleted in orthogroups with a least measured distance to each deletion model from the previous section were recorded (Table 4.5). Thus, an estimation of what functional classes of genes are more or less likely to be deleted at each branch in this *Faboideae* species tree was achieved. Terms like ‘nucleic acid binding’ and ‘recognition of pollen’ were significantly associated with orthogroups that matched varying models of deletion. Additionally, the GO terms that were most significantly associated with the neutral or ‘full retention’ model were noted separately, as this model does not model any deletions and scarce few families matched the “full retention” model the best. The significantly enriched or depleted GO terms associated with gene families that suggested a history of gene retention included functions divergent from those enriched/depleted in the deletion models, with terms such as ‘Sulfate transport’ and ‘Response to heat’ significantly under-represented in these orthogroups.

Discussion

Along with generating a large, comprehensive set of gene families for major, well-characterized crop legume genomes, this study also attempts to define a straightforward, model-based method for examining the history of duplicate genes along a phylogeny, with a focus on deletions of gene duplicates. Most of the genes in these genomes were part of some kind of duplicated gene family, though many have apparently reverted to single copies, with 868

families having exactly one gene in each species. The Orthofinder algorithm employed here appears to have been very sensitive with more than 25,000 gene families identified, of which > 19,000 were reconcilable into gene trees. Previous studies have noted that *G. max* appears to have a propensity for retention of duplicate genes even when taking into consideration its more recent mesopolyploidy event from about 8 Mya. Indeed, with 45,550 genes from *G. max* represented in gene families, by raw count it had the most genes within a gene family. Interestingly, however, it was neither the species in this study with the highest proportion of genes contained in gene families nor the species that appeared most frequently within these gene families. Instead, *P. vulgaris* had the highest percentage of genes in a gene family (93.7% of genes in an orthogroup, Table 4.1) and *A. ipaensis* appeared the most often within gene families (80.2% of families). The raw number of genes from *G. max*, then, may be a simple artifact of the large number of genes in its genome overall, at 56,044 in this study. The case of *M. truncatula* may offer some counterevidence to this however, since its gene count is also similarly high at 50,894, yet only 67.7% of those genes were placed into gene families.

Interestingly, 2.4% of *M. truncatula* genes were in species-specific orthogroups, the highest percentage among these species. This is possibly because *M. truncatula*'s genome annotation may contain many transposons or repetitive elements inaccurately classified as genes, has many genes that are incomplete models, or because it has a large amount of tandem or dispersed duplications. This underscores the importance of high-quality genome annotation in testing large-scale evolutionary hypotheses in any species, as any inferences are only as powerful as the accuracy and reliability of the annotation. Instructively, some of the largest gene families in this study were comprised of genes almost entirely from more newly created and less manually curated genome annotations like the *Arachis spp*. Presented here. For example, the

largest gene family identified had 1845 total genes, with 604, 219, 260 and 742 genes each from *A. duranensis, A. hypogaea* A and B subgenomes, and *A. ipaensis*, respectively. A simple BLASTp search of these genes against repbase showed that most or all of these genes were like transposon-related, indicating that they might not be 'true' genes with functional protein products. Comprehensive expression data for these species would be needed to verify whether these are true genes or artifacts of a highly sensitive *ab initio* gene annotation method.

While this study is focused on the broader picture of deletion and duplication in legumes, examining some of the individual gene families it identified shows that these results can be used for further study. For instance, although the gene family in Fig. 4.2 was chosen at random to represent a typical gene family from this study, it happened to include the important protein leghemoglobin A, which is known to be involved in the nodulation processes characteristic of the legumes. Interestingly, this family shows a novel duplication in the *Arachis* lineage but not the other lineages in the tree, indicating that perhaps this important nodulation-related gene has undergone functional diversification in these diploid and tetraploid peanut species. Indeed, previous studies have indicated that different forms of leghemoglobin in pea, soybean, and peanut have distinct spatial and temporal expression profiles (Hargrove et al., 1997; Kawashima et al., 2001; Lee and Verma, 1984; Marcker et al., 1984). Furthermore, it has been noted that different closely related peanut species form nitrogen-fixing nodules with different rhizobia species (Andrews and Andrews, 2017). It could be the case, then, that duplication was responsible for not only diversification of function in leghemoglobin in legumes but also for the diversification of symbiotic species recruiting in peanut owing to its peanut-specific duplication. The family presented in Fig. 4.2 represents an excellent case study that demonstrates the power of this gene family approach and the ubiquity and impact of duplication on legume genomes.

Estimating probabilities of deletion across the Faboideae using a phylogenetically-informed approach

As demonstrated in Fig 4.2, many of these legume gene families have had multiple duplication and deletion events even outside of the generally accepted time frames of the major WGD events in the legume lineage. Assessing the extent of gene duplication and deletion along with WGD events is a complex issue, and has been approached in many ways in other studies (Maere et al., 2005; Schmutz et al., 2010b; Schnable et al., 2012; Schnable et al., 2011; Wang et al., 2017a; Wang et al., 2011). In this case, a simple phylogenetic comparison method was chosen, as it was feasible for the large number of gene families among these legumes. Furthermore, since deletions of genes are likely the most drastic outcome of gene duplication (though sometimes, deletions are presumed to be mitigated by dosage compensation) (Gu et al., 2003), deletions were chosen as the focus of the modeling for this study as opposed to recurring duplications, or any other model. Notably, the method employed here using phylogenetic tree comparisons means that these models do not necessarily discriminate between stochastic or 'background' loss of genes and the typically rapid loss of duplicates directly after a WGD. Regardless, this approach still showed that gene loss following the ~60 My old *Faboideae* WGD event was rapid and common, with this model being both the model with the lowest average distance to the observed gene trees (Fig 4.4). Both sides of the *Glycine* branch, either before/concurrent with the *Glycine* WGD (Fig 4.4) or after the WGD had relatively lower estimated probabilities of deletion, as these models were calculated to have a large average distance from the data overall or matched few gene families better than any other model (Fig 4.5). The rates of deletions on the branches lacking WGDs were appreciably lower than those with WGDs, in contrast. The *Arachis* lineage in particular had notably low estimated rates of

deletion, even with the very short branch lengths in that clade taken into consideration. The branch at the base of the *Arachis* clade did have a slightly higher average deletion probability, along with the branch representing the common ancestor of *Medicago, Phaseolus,* and *Glycine*. The *Medicago* branch showed a curiously high deletion likelihood, but this may also be an artifact of misannotation of transposable elements as genes, as discussed earlier regarding *Medicago*'s high amount of species-specific gene families. Transposable elements are likely to be spuriously included in gene families that show little relation to the rest of the evolutionary topology of the rest of the legumes, to be dynamically gained and lost relatively rapidly over time, and to be deleted at very high rates.

Estimating gene birth and death rates and WGD retention rates using maximum likelihood reveals significant differences in the Faboideae *and* Glycine *WGDs*

While a straightforward phylogenetic method, as outlined above, can provide insights as to the probabilities of duplicate gene deletion among clades and across time, there are other methods that use maximum likelihood (MLE) or Markov chain modeling approaches. Most MLE methods of the rates of gene duplication and deletion model these as birth-death processes, which is a kind of continuous Markov process, where genes are equally likely to be duplicated or deleted across a taxonomic clade according to stochastic chance. The observed rates of gene duplication (often denoted λ) and deletion (μ) can be calculated using a birth-death model through maximum likelihood methods. Some methods, however, have allowed for varying rates of birth and death across different branches of a lineage, or have attempted to model instantaneous duplication or triplication (WGDs or WGTs) of genes along with a birth-death mode (Ghenu et al., 2016; Hahn et al., 2007). In this study, both were integrated using the WGDgc method (Rabier et al 2014), which is limited in its ability to estimate differing

background rates of birth and death but can estimate different rates of retention for various WGD events in a lineage.

The results in Table 4.3 and Figs. 4.6 and 4.7 show that background birth and death rates were in line with estimates for lineages estimated in other studies in both plants and animals (Akhunov et al., 2007; Hahn et al., 2007; Lynch and Conery, 2000). The choice of dataset used to estimate these parameters slightly affected these estimates, with the inclusion of computationally expensive larger gene families affecting results much more than the overall number of gene families sampled (Figs 4.6 & 4.7). Including gene families with 50 or fewer members rather than 20 yielded a lower estimate for the death rate but a significantly higher estimation of birth rate (Fig 4.6), as expected, since including larger families means including more duplication events. This was also true of retention rates, where including families with 50 or fewer members gave a somewhat higher retention estimate for the *Faboideae* WGD, but a negligible effect on the *Glycine* WGD retention estimate (Fig 4.7). The higher birth rate estimated when larger families (from 21-50 total gene members) were included indicates that just a few hundred large gene families dominated gene duplication outside of WGDs. These families are likely to be either TE-related (as in the case of the massive *Arachis* families discussed earlier that were likely a result of mis-annotation) or belong to some of the classes (transcription factors, protein dimerization genes, pollen recognition genes, etc.) identified in GO enrichment analysis.

Regardless, in both cases, the general patterns of higher death rates than birth rates and a much higher retention rate for the *Glycine* WGD than the *Faboideae* WGD held true. Death rates were about 3 times higher than birth rates, indicating that gene loss in these legumes was relatively common over time post-WGD, and that other duplications (tandem, dispersed, or segmental) were relatively rarer and probably concentrated in much larger gene families. The

retention rates between the two different WGD events in this lineage were very divergent, with the *Glycine* duplication retaining 2 gene copies at a far higher rate than genes from the *Faboideae* duplication (~78% vs. 20%). This suggests that these two events were under very different constraints; for instance, the *Glycine* duplication may have been a segmental allopolyploidy or autopolyploidy and the *Faboideae* duplication may have been an allopolyploidy event. Other evidence from recent studies supports this, including evidence of biased fractionation post-WGD for the *Faboideae* event and evidence against bias in the *Glycine* event (Xu et al., 2018). The reason these *Glycine* duplicate copies may be retained at a higher rate than those from the *Faboideae* event may be because there were few or no pre-existing sequence, expression, or methylation differences in the copies pre-*Glycine* WGD owing to its more autopolyploid-like nature. This in turn would mean that gene dosage balance may have constrained all genes in *Glycine* and prevented loss of any one copy of a set of duplicates except for a few deletion-tolerant functional classes of genes like TFs or regulatory network-related genes.

Enrichment of GO terms and association of GO with gene family size

In order to test whether different functional classes or broad categories of genes experience divergent evolutionary histories, all genes of all genomes considered here were associated with either a publicly available GO annotation or, in the case of tetraploid peanut, a *de novo* GO annotation. Testing for GO enrichment within orthogroups that best matched different models of deletion showed that there were significant differences in the general types of genes that were more or less likely to be deleted across different lineages in the legumes. Results from this GO enrichment analysis show that gene classes which are generally thought to participate in gene networks, where many genes interact in modulating each other's expression or the

multimerization of their resulting protein products, are significantly under or overrepresented with membership in orthogroups that experienced Nucleic acid binding genes, which are often transcription factors, were less likely to be deleted at the base of the *Arachis* species clade, but were more likely to be deleted in the common ancestor of *Medicago*, *Phaseolus,* and *Glycine*, or in either of the individual *Phaseolus vulgaris* or *Glycine max* species branches (Table 4.4). Protein dimerization genes, which like DNA-binding genes often participate in network formation and regulation, were also observed to be significantly enriched or depleted in varying deletion events. Like DNA-binding genes, protein dimerization genes were underrepresented in deletions at the base of the *Arachis* clade, but overrepresented in deletions in the common ancestor of *Medicago, Phaseolus* and *Glycine*. Unlike the nucleic acid binding genes, however, these were underrepresented in deletions in the *Phaseolus* lineage.

Previous studies have noted that not all classes and functions of genes are duplicated, deleted, or retained post-WGD equally. In particular, it has been noted that DNA-binding transcription factor (TF) genes are often duplicated and are generally in large gene families (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004). A series of ANOVA tests for each GO term against the size of the gene families that each GO-related gene was included in revealed that most terms were negatively correlated with gene family size. This could mean that most identifiable functional classes of genes tend toward diploidization or deletion of extra copies. In contrast, a few select functional categories not only tolerate duplication but also perhaps benefit organismal fitness by allowing for new beneficial functions or network connections to be made. This is supported by the strong positive correlation with gene family size for genes with annotated functions including nucleic acid binding, zinc ion binding (often a part of zinc-finger containing genes, which bind to nucleic acid), protein dimerization (involved in creating protein

multimers and network interactions), and response to auxin stimulus (which often involves pathways that result in transcriptional regulation) (Quint and Gray, 2006).

In summary, these results (Table 4.5) suggest that genes that participate in complex regulatory networks with protein-protein and protein-DNA interactions had more dynamic histories of duplication and deletion. The gene balance or dosage balance hypothesis predicts that when members of a network or multimeric complex are duplicated, the stoichiometric ratios of the interactions of the members help to maintain these genes in duplicate (Birchler and Veitia, 2007; Birchler and Veitia, 2014; Tasdighian et al., 2017; Teufel et al., 2016). For example, if an enzyme requires 2 proteins of gene "A" and 1 of gene "B" to function, duplicating both genes would result in an equal ratio of both products, yielding a functional product. However, if the dose of the product relative to the cellular component it acts upon (for instance, a transcription factor on its target gene) is not as crucial to its function or overall organismal fitness, extra copies of these genes may allow for new functions to arise or network connections to be made. In this way, the "Xfunctionalization" hypothesis predicts that in some cases, extra copies of genes relax selection on some copies and allow for functional and network diversity (Pastor-Satorras et al., 2003; Wagner, 1994, 2001). Furthermore, dosage compensation might account for those cases where deletions in TFs or network-related genes have occurred in large numbers. Dosage compensation posits that in some cases, gene deletions, nonfunctionalization, or pseudogenizations have little phenotypic effect. This could be due to either a duplicate gene providing the necessary protein product, or other members of the regulatory network or alternative pathways leading to the same metabolic outcome without the deleted gene (Gibson and Spring, 1998; Gu et al., 2003; Holstege et al., 1998; Maslov and Sneppen, 2002; Nowak et al., 1997). In many of these cases where dosage compensation is hypothesized to be at play, it

has been noted that often the higher-expressed gene is less likely to be deleted, perhaps suggesting that this has a larger phenotypic effect than deleting a lesser-expressed copy. It is possible, then, that in different lineages and at different points in time among these legumes, one or several of these mechanisms predominated in the evolution of the regulatory networks that shape these genomes. While most of the species considered in this study are domesticated crops, these processes have apparently been at play for millions of years, and thus are probably not associated with domestication or improvement of these plants over the last few thousand years, but rather their evolution and adaptation over millions of years.

Conclusion

The crucial role of gene and genome duplication in the evolution and diversification of not only flowering plants but also all life on earth is more and more appreciated as more genomes are sequenced, assembled, and analyzed. Selection, genetic drift, gene flow, and point mutations are all methods by which allele frequencies can shift or fix within a population, or by which new variants can arise, but few evolutionary forces have as drastic an impact on genetic variation as duplication. The presence of multiple gene copies not only often has phenotypic effects of its own, the spare DNA for evolutionary forces to act upon can give rise to a plethora of new phenotypic variants for a species, or can even give rise to new species altogether. While there is evidence of ancient whole-genome duplications across the tree of life, there are many other duplication events within genomes that occur regularly. Segmental, tandem, and dispersed duplications are also important processes by which new genes and gene variants appear.

As such, genes are typically part of duplicated gene families with many copies of genes present across a genome and between species, with considerable variation in gene family size, membership, and topology. In this study, these characteristics of gene families were examined in

select well-characterized *Faboideae* (Papilionoid) legume genomes and grapevine, revealing a dynamic history of duplication and deletion among these genomes. While the expansion of gene families post-WGD was apparent across the data, gene losses were quite common, even if they were probably not equally likely across lineages. Gene losses were concentrated in a few branches along the species tree of these genomes, especially following the WGD event in the MRCA of the *Faboideae*, though the background rate of gene deletion was relatively high throughout the family. One notable exception could be noted in the *Glycine* lineage, where the WGD in the ancestor of this genus was estimated to have retained far more of its genes proportionally than the *Faboideae* WGD. This is in line with growing evidence that the *Glycine* WGD about 13 Mya was more like an autopolyploidy than originally assumed. Therefore, the often-noted high copy number of genes in soybean may be primarily due to the high retention rate of duplicates following the *Glycine* WGD rather than an especially high background gene birth rate or a low background gene death rate. In any case, soybean appears to be something of a unique case among the legumes in this regard, and further study into why it has apparently retained so many duplicates, especially from its lineage-specific WGD, is worth investigating.

These findings have many implications for the evolutionary history of legumes and for their future breeding efforts. For example, with more gene copies for any given gene in soybean, it may be that the phenotype of the plant is more plastic than it would be with fewer copies, since the remaining functional copy could mask an otherwise lethal variant in one copy of a gene. Alternatively, this could mean that since gene copies are maintained in soybean in a dosage balance, changing the expression or function of one copy could affect the function of its sister copy. Experimental studies modifying expression or function of gene copies, or studies examining how different copies of genes in soybean among different genotypes can affect

phenotype are warranted by this finding. Since legumes are not only one of the largest plant families but also include many of the most valuable crops on earth, a thorough knowledge of how multiple copies of genes contribute to the phenotypes of these crops.

Table 4.1. Basic statistics describing the orthogroups. A "species-specific" orthogroup is an orthogroup containing only paralogs or containing only genes from one species. Note that these do *not* include putative orphan genes.

	Arachis duranensis	***Arachis hypogaea-A***	***Arachis hypogaea-B***	***Arachis ipaensis***	***Glycine max***	***Medicago truncatula***	***Phaseolus vulgaris***	***Vitis vinifera***
Number of genes	36734	28671	32132	41840	56044	50894	27433	26346
Number of genes in orthogroups	33041	26830	29807	36488	45550	34461	25716	18621
Percentage of genes in orthogroups	89.9%	93.6%	92.8%	87.2%	81.3%	67.7%	93.7%	70.7%
Number of orthogroups containing species	19022	16924	18233	20159	16938	16000	16040	12978
Percentage of orthogroups containing species	75.6%	67.3%	72.5%	80.2%	67.4%	63.6%	63.8%	51.6%
Number of species-specific orthogroups	10	0	1	7	18	79	7	34
Number of genes in species-specific orthogroups	32	0	2	24	118	1228	34	315
Percentage of genes in species-specific orthogroups	0.1%	0%	0%	0.1%	0.2%	2.4%	0.1%	1.2%

Table 4.2. Deletion model statistics. "Mean treeKO distance" refers to the mean distance between the deletion model and all orthogroups in the study. "Branch length" is the newick tree branch length on the species tree for the deletion model. The "corrected mean" is the treeKO distance divided by the branch length to control for evolutionary time. The "RGB value" is the hex coded RGB value for the branch in Fig. 4.4 calculated by [Red = (corrected mean), Blue = 1-(corrected mean), Green = 0].

Model	**Mean treeKO distance**	**Branch length**	**Corrected mean**	**RGB value**
A	0.2370353	0.181786	0.76691521	#C4003BFF
B	0.2595158	0.163729	0.63090197	#A1005EFF
C	0.3010029	0.0250404	0.08318989	#1500EAFF
D	0.3008046	0.0273659	0.09097567	#1700E8FF
E	0.3016952	0.0318644	0.10561786	#1B00E4FF
F	0.3016952	0.0338909	0.1123349	#1D00E2FF
G	0.3016018	0.0341298	0.1131618	#1D00E2FF
H	0.3016018	0.029235	0.09693245	#1900E6FF
I	0.2667403	0.177655	0.66602234	#AA0055FF
J	0.2467189	0.177655	0.72007057	#B80047FF
K	0.2665277	0.101948	0.3825044	#62009DFF
L	0.2909101	0.101948	0.35044504	#5900A6FF
M	0.2930073	0.0813758	0.27772614	#4700B8FF
N	0.3015352	0.0813758	0.26987166	#4500BAFF

Table 4.3. Parameters of legume orthogroup evolution for various data subsets determined using the modified WGDgc nonlinear optimization method. The substitution rate per million years is also included for reference, as this was used to calibrate all values in the table. Thus, all values in this table represent their value per million years, with the exception of the Log Likelihood. "20 and under" and "50 and under" refers to the number of genes in the orthogroups included – the former being a subset including only orthogroups with 20 genes or fewer, and the latter being a subset including only orthogroups with 50 genes or fewer. The p(1) and p(2) parameters for the WGD events represent the probability that 1 or 2 genes are retained from the duplication respectively.

Dataset	Log likelihood	Birth rate	Death rate	Faboideae WGD p(1)	Faboideae WGD p(2)	Glycine WGD p(1)	Glycine WGD p(2)
Families 20 and under	-199031.9878	0.003676407	0.01176658	0.728581148	0.271418852	0.206850772	0.793149228
5000 families 20 and under	-41765.54492	0.003481959	0.012156171	0.665221668	0.334778332	0.212796739	0.787203261
Families 50 and under	-229220.8767	0.005288131	0.011418608	0.801697898	0.198302102	0.223586069	0.776413931
5000 families 50 and under	-45790.10942	0.005258508	0.011287458	0.81372274	0.18627726	0.215288285	0.784711715
Substitution rate per My	0.003322905						

Table 4.4. GO terms most strongly associated with increased or decreased orthogroup size, determined via a separate one-way ANOVA for each GO term. Positive values indicate GO terms that are strongly associated with larger orthogroups (more genes), and negative values indicate terms strongly associated with smaller orthogroups (fewer genes). P-values were

GO term	Description	Coefficient	P-value (corrected)
GO:0008270	zinc ion binding	237.777519	0
GO:0046983	protein dimerization activity	267.667113	0
GO:0008234	cysteine-type peptidase activity	287.035753	1.42E-177
GO:0003676	nucleic acid binding	110.337616	2.82E-146
GO:0005515	protein binding	-39.23305	1.75E-61
GO:0055114	oxidation-reduction process	-51.303517	1.49E-47
GO:0005524	ATP binding	-40.237825	3.29E-45
GO:0003824	catalytic activity	-55.661267	9.17E-42
GO:0016020	membrane	-49.052534	3.44E-36
GO:0048544	recognition of pollen	167.451866	1.02E-31
GO:0016021	integral to membrane	-50.858896	1.28E-30
GO:0006508	proteolysis	67.9901978	2.21E-30
GO:0008152	metabolic process	-51.466843	2.93E-28
GO:0016491	oxidoreductase activity	-52.639296	1.37E-27
GO:0005634	nucleus	-57.637483	1.79E-25
GO:0055085	transmembrane transport	-52.000816	1.22E-23
GO:0009733	response to auxin stimulus	196.227659	7.82E-23
GO:0003700	sequence-specific DNA binding transcription factor activity	-58.565765	8.89E-21
GO:0006278	RNA-dependent DNA replication	253.885746	7.23E-19

Bonferroni corrected.

Table 4.5. GO terms most strongly associated with each deletion model. Non-significant GO terms are not included, with a maximum of the top 5. P-values were corrected for multiple comparisons with a Bonferroni correction (2282 comparisons). Odds ratios above 1 indicate overrepresentation or enrichment in orthogroups most closely matching each deletion model; ratios below 1 indicate depletion or underrepresentation.

Deletion Model	GO term	Description	P-value (corrected)	Odds ratio
A	GO:0008234	cysteine-type peptidase activity	4.81E-86	5.86431426
A	GO:0006508	proteolysis	1.36E-26	1.81039799
A	GO:0000786	nucleosome	2.73E-19	0.01532915
A	GO:0015935	small ribosomal subunit	1.30E-17	0.03108903
A	GO:0003735	structural constituent of ribosome	3.01E-17	0.52641735
B	GO:0008270	zinc ion binding	5.47E-23	0.44761718
B	GO:0003676	nucleic acid binding	1.30E-21	0.42359209
B	GO:0046983	protein dimerization activity	4.44E-15	0.32205502
B	GO:0010181	FMN binding	1.23E-12	5.9647896
B	GO:0009664	plant-type cell wall organization	6.00E-09	5.62895546
D	GO:0016787	hydrolase activity	1.97E-20	137.012903
D	GO:0008152	metabolic process	4.58E-15	55.6437099
I	GO:0006278	RNA-dependent DNA replication	1.03E-51	67.4149804
I	GO:0003964	RNA-directed DNA polymerase activity	2.11E-50	58.3142801
I	GO:0046983	protein dimerization activity	1.74E-16	2.22181452
I	GO:0003676	nucleic acid binding	1.68E-15	1.78364062
I	GO:0010333	terpene synthase activity	1.11E-13	6.99583665
J	GO:0043531	ADP binding	1.13E-46	4.76927372
J	GO:0009733	response to auxin stimulus	2.59E-43	0.0690139
J	GO:0008234	cysteine-type peptidase activity	6.80E-37	7.98241295
J	GO:0005634	nucleus	1.70E-31	0.5418569
J	GO:0006952	defense response	3.16E-15	2.57667525
K	GO:0008270	zinc ion binding	1.94E-172	3.58541666
K	GO:0005840	ribosome	1.80E-78	4.12601302
K	GO:0006412	translation	4.44E-76	3.91759462
K	GO:0003735	structural constituent of ribosome	3.81E-74	3.8931936
K	GO:0009579	thylakoid	3.85E-69	139.898398
L	GO:0046983	protein dimerization activity	9.10E-102	0.22435489
L	GO:0009607	response to biotic stimulus	7.32E-68	0.04525692

L	GO:0008270	zinc ion binding	1.33E-32	3.03180169
L	GO:0008152	metabolic process	9.39E-21	2.83495212
L	GO:0003676	nucleic acid binding	2.37E-19	2.42229383
M	GO:0030247	polysaccharide binding	1.37E-33	17.4108376
M	GO:0043531	ADP binding	1.27E-20	4.79635898
M	GO:0008408	3'-5' exonuclease activity	7.66E-19	30.5628665
M	GO:0046982	protein heterodimerization activity	1.32E-16	10.1562979
M	GO:0000786	nucleosome	9.90E-13	10.485166
N	GO:0003676	nucleic acid binding	6.13E-67	3.38733159
N	GO:0048544	recognition of pollen	1.51E-64	0.11166125
N	GO:0006468	protein phosphorylation	4.59E-51	0.54292344
N	GO:0004672	protein kinase activity	1.20E-50	0.54388949
N	GO:0016760	cellulose synthase (UDP-forming) activity	9.84E-46	0.0613063
O	GO:0015116	sulfate transmembrane transporter activity	1.52E-20	228.378389
O	GO:0008272	sulfate transport	1.52E-20	228.378389
O	GO:0008271	secondary active sulfate transmembrane transporter activity	9.87E-15	215.312011
O	GO:0016787	hydrolase activity	4.66E-08	11.4529679
O	GO:0009408	response to heat	8.72E-03	41.7294379

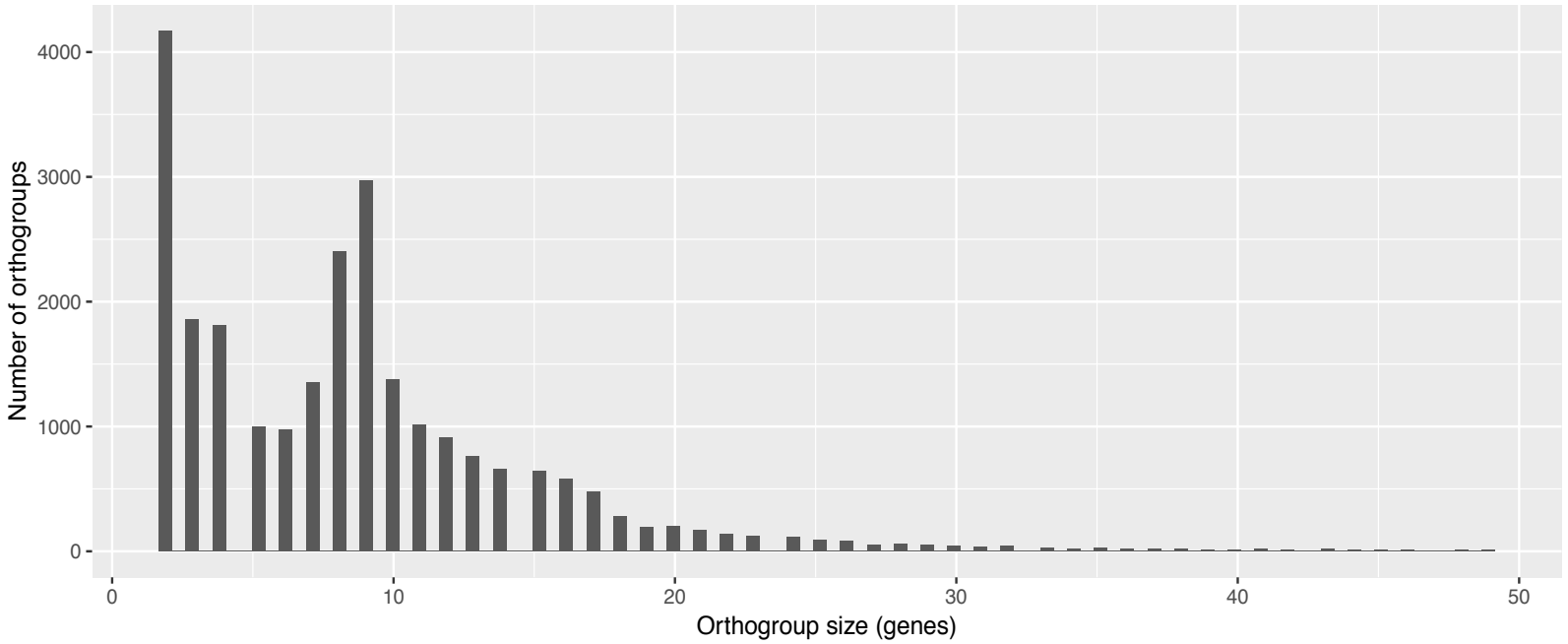

Figure 4.1. Distribution of gene families identified via Orthofinder. 238 gene families larger than 50 genes and all single-gene gene families are excluded. The peak at n=1 genes represents an abundance of putative orphan genes, which had no identifiable orthologs or paralogs among the species set considered here.

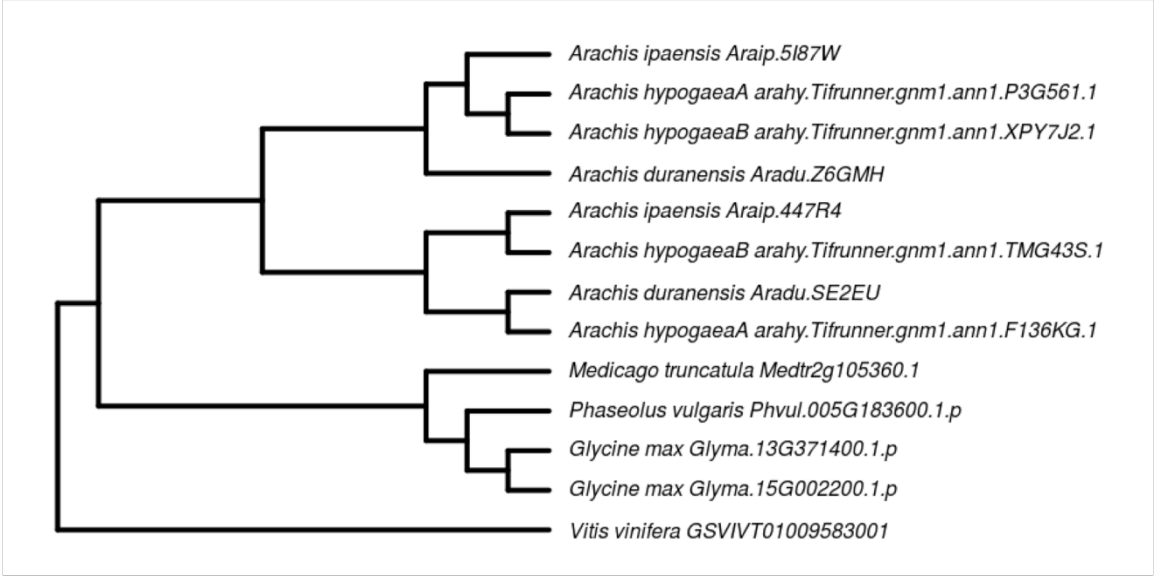

Figure 4.2. Randomly selected example reconciled gene family tree from Orthofinder. This gene family has putatively experienced a deletion at the base of the *Faboideae* post-WGD, a retained duplication post-*Glycine* WGD, and a non-WGD duplication at the base of the *Arachis* clade. A UniProt search and GO analysis suggests this family is the Leghemoglobin A family, with genes related to e.g. oxidation-reduction processes (GO:0055114), and with genes that interact with heme groups (GO:0020037). This family, then, is important for the legumes, as leghemoglobin is important in the symbiotic relationship between legume roots and the nitrogen-fixing rhizobia that characterize the unique, valuable characteristics that define legumes. The expansion in the *Arachis* lineage suggests a unique diversification of leghemoglobin and rhizobial interactions in that clade.

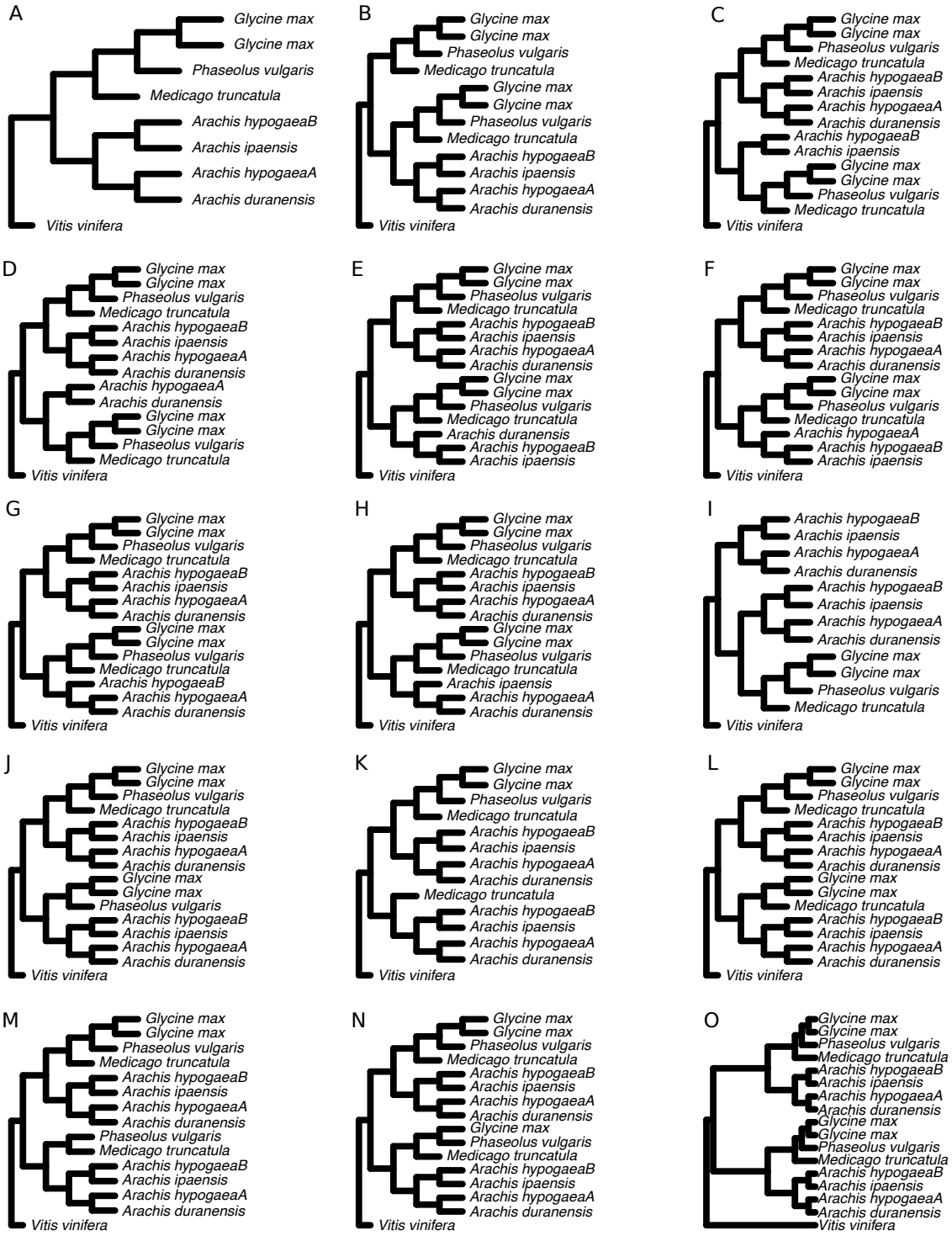

Figure 4.3. Models of deletion used in the phylogenetic deletion probability estimation method. Each model represents a single deletion in one branch of the species tree. Letters match the branch designations in Figure 4.4. **A:** a deletion in the common ancestor of the *Faboideae*

post-WGD; **B:** a deletion in the base of the *Arachis* clade in one copy post-WGD; **C:** a deletion in the common ancestor of *A. duranensis* and the peanut A subgenome; **D:** a deletion in the common ancestor of *A. ipaensis* and the peanut B subgenome; **E:** a single deletion in the A subgenome of peanut; **F:** a single deletion in *Arachis duranensis*; **G:** a single deletion in *Arachis ipaensis*; **H:** a single deletion in the peanut B subgenome; **I:** a deletion in the MRCA of *Medicago, Phaseolus* and *Glycine*; **J:** a single deletion in *Medicago*; **K:** a deletion in the common ancestor of *Glycine* and *Phaseolus*; **L:** a single deletion in *Phaseolus*; **M:** a deletion in *Glycine* pre-WGD; **N:** a single deletion in *Glycine* post-WGD; **O:** no deletions in any branch (full retention).

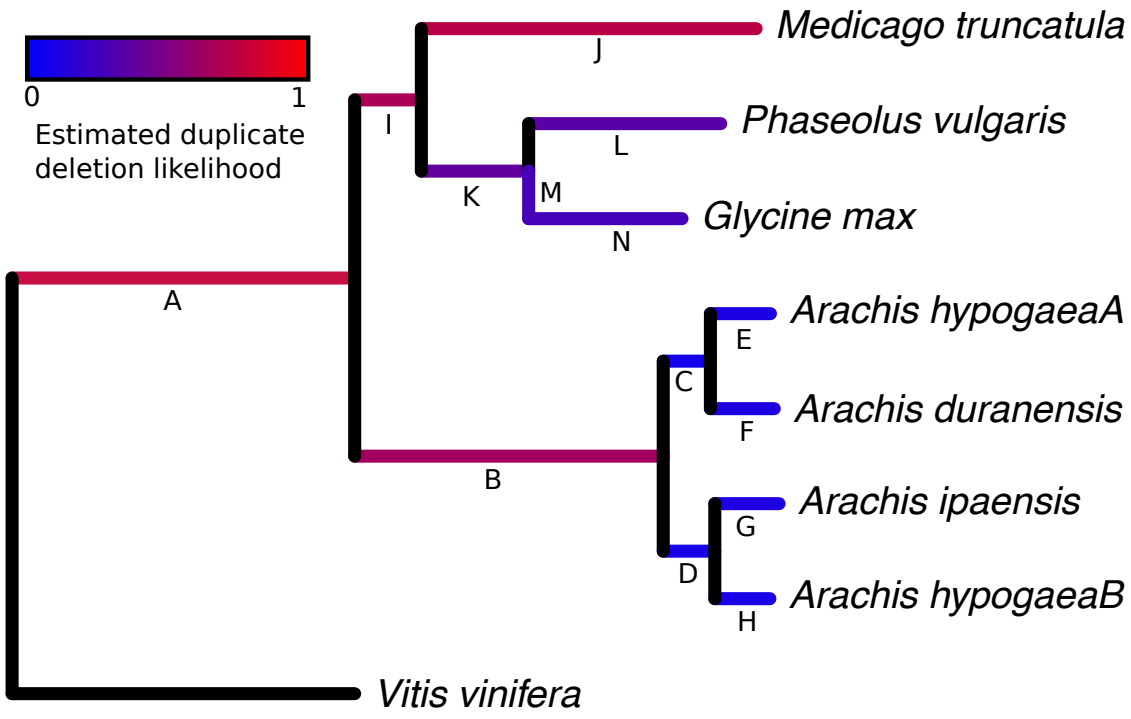

Figure 4.4. Heat-mapped species tree showing the estimated probability of deletion determined through the phylogenetic comparison method. A redder color indicates higher likelihood of deletion. Values represent those shown in Table 4.2. Since the treeKO algorithm outputs numbers from 0 to 1, the scales of the colors are constrained in this range. Final color values are the average treeKO distance of the branch deletion model to all gene families divided by the estimated species tree branch length for that branch.

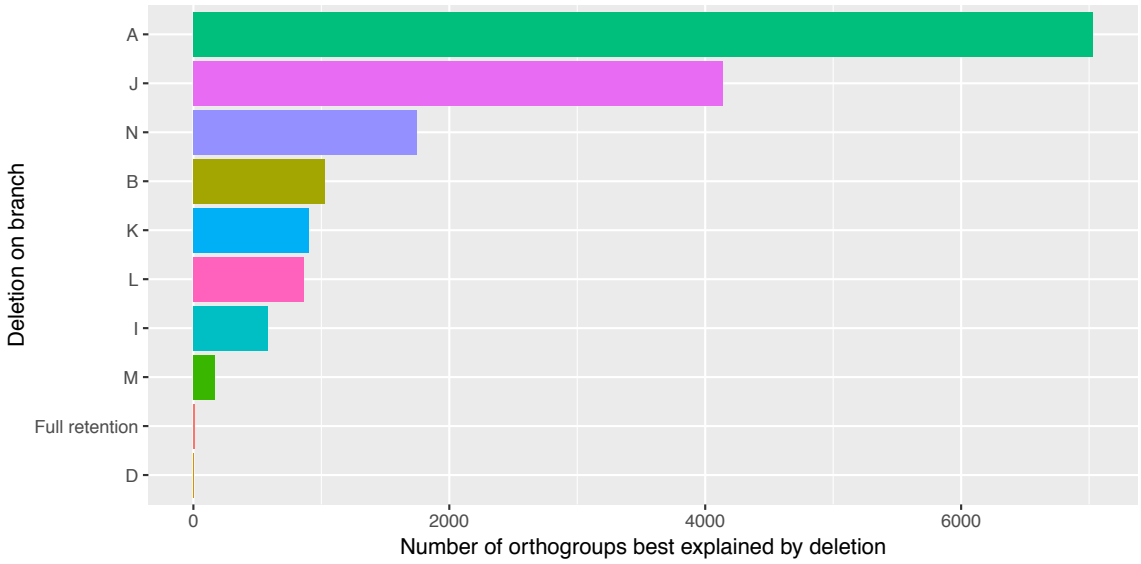

Figure 4.5. Number of gene families with the least distance to each deletion model from Fig 4.3 and 4.4. For each gene family, the single model tree with the least calculated treeKO distance to the gene tree was chosen as the "best fit" model for that group. Only one family matched C or D equally well, and none matched E, F, G, or H best, and thus those models are not included.

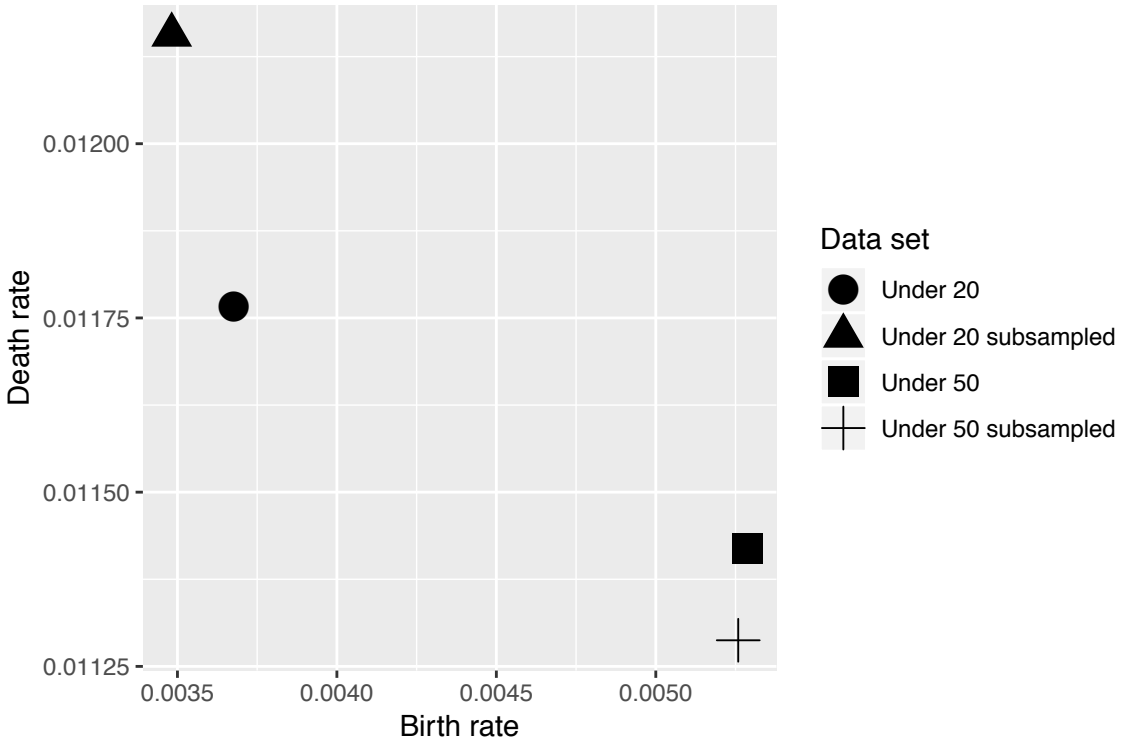

Figure 4.6. Maximum likelihood derived parameters for the overall gene birth and death rate for the gene families identified by Orthofinder in this study. The birth rate refers to the number of duplications per gene per million years in all genomes. The death rate is the expected number of duplicate gene deletions per gene per million years. The rates were calibrated using a molecular clock calculation from the 'ape' R package's chronoMPL function. The datasets refer to either subsets of gene families with ≤20 or ≤50 total members, and 'subsamples' of 5000 randomly chosen families from those ≤20- or ≤50-member families. The range of the axes is notable, with far more variation in the estimated birth rate than death rate.

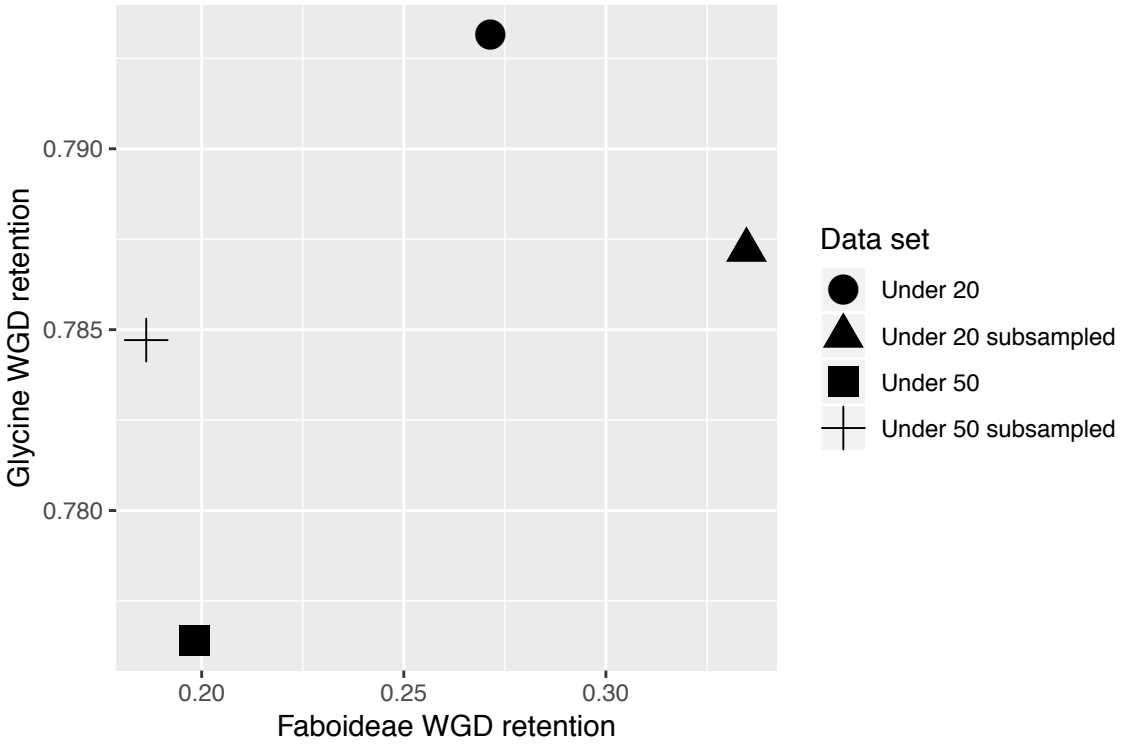

Figure 4.7. Variation in the estimated probabilities of gene retention for the two major WGDs in legumes (x and y axes) for each of the 4 datasets described in Fig 4.6. The retention number is expressed as a probability from 0 to 1 that a given gene is retained in two copies immediately post-WGD. These parameters do *not* model the probability that a duplicate arising from a WGD is deleted millions of years after the WGD - these kinds of deletions are instead accounted for by the background 'death rate' in Fig 4.6.

CHAPTER 5

DOMESTICATION SWEEPS IN DUPLICATED AND DISPENSABLE GENES IN SOYBEAN

Introduction

Modern agriculture is dependent on improved crop varieties that have been shaped by human selection over thousands of years (Smith and Nesbitt, 1995). Starting with the first domestication of cereals in the Middle East over 10,000 years ago, humans have imprinted the indelible signatures of artificial selection upon the genomes of the plants cultivated around the world (Doebley et al., 2006; Poncet et al., 2004). Traits like absence of seed shattering, prevention of lodging, loss of dormancy, and larger and more numerous seeds mark the 'domestication syndrome' common to many crop plants selected by humans for cultivation (Hammer, 1984). The domestication syndrome has been noted in wheat, rice, barley, triticale, soybean, common bean, maize, millet, sorghum, sunflower, and more (Koinange et al., 1996; Milla et al., 2015; Sakuma et al., 2011; Sedivy et al., 2017; Wright et al., 2005). Thus, given enough generations between a wild plant and its domesticated counterpart and selection for domestication traits, significant changes in morphology, development, and genetics can be seen.

Domestication involves selecting only a few individuals with characteristics suitable for cultivation to advance to the next generation. This imposes a genetic bottleneck upon the plant genome, wherein the genetic diversity of the populations selected for advancement is progressively lessened as more and more selections are made (Eyre-Walker et al., 1998). This results in an improved variety with stable, predictable growth and yield, a more bountiful

harvest, seeds that stay on the plant instead of shattering and spreading, and many more valuable traits. However, during domestication, much of the genetic diversity originally present in the wild plant populations is lost (Hyten et al., 2006). Thus, the selection of plants for valuable domestication traits leaves identifiable marks upon the genomes of the crop plants of today, allowing geneticists to untangle the history of these plants.

Artificial selection narrows the genetic diversity of a plant as just as it confines its diversity in morphology or development. Thus, by studying how different segments of a plant genome have been reduced in diversity or differentially selected from the original wild varieties of antiquity to the elite cultivated varieties grown today, it is possible to identify what genes or genomic sequences are potentially responsible for domestication traits (Myles et al., 2011; Wang et al., 2017b). While the inheritance of traits and diversity of alleles for those traits have been studied in crop plants for many years, widely available and inexpensive sequence data has only recently allowed for studying the population genetics of these traits in finer detail than ever before. The simplest method for identifying these regions using sequence data is simply to compare the basic nucleotide diversity (π) within a genome segment (e.g. a gene) and identify regions where this diversity has been greatly reduced between wild and elite lines (Dagan et al., 2007; Vida, 1978). Other more sophisticated statistical measures like Ka/Ks, Tajima's D, F_{ST} statistics and more can give further insight as to how the genetics of a region have changed through domestication and how specifically (positive, negative, balancing, etc.) selection has acted upon them (Hartl et al., 1997; Korneliussen et al., 2013; Weir and Cockerham, 1984).

Identifying genomic regions involved in domestication and improvement is an active area of research in crop plants, as it not only improves understanding of how crop plants came to be, but also helps point to possible routes for improving crop plants further through breeding.

Classical and contemporary genetic techniques like biparental mapping or transformation have allowed for impressive leaps in the understanding and application of how crop plants have been selected for their valuable traits over many generations of human intervention (Moose and Mumm, 2008; Visarada et al., 2009). These advancements in understanding have historically been primarily focused on comparing homologous segments within genomes; for example, SNPs in important genes that control flowering time, haplotype blocks in large genome segments that are associated with domestication and improvement, or inserting via transformation different forms of candidate genes from different genotypes and species (Li et al., 2009; Vanblaere et al., 2011). By and large, studies investigating these traits have focused on variants of an allelic nature. However, less is known about how domestication and crop improvement might be result from large-scale structural variation in genomes, such as whole genome duplications, polyploidy, tandem duplications, copy number variations, and presence-absence variation (Morrell et al., 2012). Some traits have been identified that are associated with structural variants, like photoperiod sensitivity in wheat or fruit shape in tomato (Nishida et al., 2013; Rodriguez et al., 2011). While cytogenetic studies in crop plants established long ago that chromosomes are highly dynamic cell components, the breadth of variation in structure was largely underappreciated until recently with long-range sequencing techniques. This is largely because *de novo* assembly of a genome at the chromosome scale has been a daunting undertaking, requiring immense resources and knowledge to achieve. Thus, high-quality genome assembly was confined to one (often inbred, diploid) reference genotype for each species. While this made genome assembly more feasible, its inherent drawback is that it collapses both the allelic and structural diversity of the species into one representative genome sequence (Pinosio et al., 2016; Vernikos et al., 2015).

With even more inexpensive sequencing technology available, newer studies have doubled back on genome assembly, looking to assemble high-quality chromosome-scale genomes for many accessions/genotypes within a species. Crop plants are of great interest in these efforts, as domestication bottlenecks have not only reduced allelic diversity in genes and congenic regions across crop genomes but have also reduced their structural diversity. Entirely new genes or chromosome segments are found in wild or unimproved germplasm, and these may harbor valuable traits (Li et al., 2014). While genetic mapping can in some cases reveal traits associated with structural variants by finding markers associated them, the genes or sequences contained within the structural variants or the exact deletions, duplications, inversions, or other changes that define them cannot be captured in this way.

Capturing non-SNP variation with sequencing data, even with improved short read sequencing technology, remains a difficult proposition. Structural variants are often contained within or defined by repetitive elements, which are notably difficult or impossible to resolve with short-read sequencing. In essence, if a variant (whether it be a copy number variant, a large or complex indel or inversion) is considerably longer than the average insert length or sequence read of a sequencing run, the variant can be missed during assembly or alignment to a reference genome. In the latter case, aligning to a reference genome hampers the resolution of the true diversity in resequencing studies (Alkan et al., 2011). Any sequence that is not directly alignable to the reference genome will be discarded by most alignment algorithms, leading to many potential "true" variants being missed by these (Mielczarek and Szyda, 2016). Long read technologies and optical mapping can help resolve some of these structural variants, but they remain cost-prohibitive for extensive resequencing of germplasm for diversity studies (Liu et al., 2012; Schadt et al., 2010). Regardless, some structural variants can be detected with certain

methodologies using short-read data. Simple presence-absence variation can be detected via mapping resequenced accessions to a reference genome, ensuring unique mappings, and looking for genomic segments which lack coverage from these mappings (Tan et al., 2012) and by de novo assembly of unaligned sequence reads. Other more complex methods offered by certain software suites can detect some larger SVs using short read data as well (Stütz et al., 2012).

While studies of domestication of crop plants using SNP data or structural variants has been performed for many crops in the past, little is known about how domestication is affected by the highly duplicated nature of plant genomes. All sequenced crop genomes show an extensive, long-reaching history of polyploidy and gene duplication, resulting in modern plant genomes with many duplicate genes and large gene families (Panchy et al., 2016). While this has sparked investigation into the evolutionary history of plants (i.e. processes that take millions of years), comparatively less attention has been drawn to how duplication in plant genomes may have played a role in domestication or improvement of these plants (processes that take thousands of years or many generations). While genes responsible for domestication or improved agronomic traits have been identified (Gross and Olsen, 2010; He et al., 2011), less is known about how duplicated copies of these agronomically critical genes or genomic regions might have been affected by their duplicated nature. For instance, if a gene is identified as controlling seed shattering, do its duplicate copies in the genome also show signatures of selection via domestication, and, if so, how does this affect the genetic architecture of domestication traits?

With the recent release of extensive resequencing data for soybean (*Glycine max* L.) and its close wild progenitor, *Glycine soja*, an examination of the extent of not only domestication sweeps and selection during improvement but structural variation in germplasm of this

economically crucial crop is possible (Valliyodan et al., 2016). Soybean is a large, highly duplicated genome, with over 55,000 identified genes and at least 3 detectable whole genome duplication (WGD) events in its genome at ~130, 55, and 8-13 Mya respectively (all of which are shared with *G. soja* and other *Glycine* species) (Schmutz et al., 2010b; Zheng et al., 2013). Soybean also shows many duplicate gene copies with similar or identical expression profiles, suggesting that duplicate genes in this genome often have overlapping functions (see Chapter 2). This study aims to determine whether duplicate gene copies in soybean were selected separately or together during domestication, whether genes in duplicated families are more or less likely to be involved in domestication, and how both of these interact with genes that are present or absent in various soybean accessions.

Resequencing and variant call data for 480 soybean lines were used to investigate how domestication sweeps have affected different classes of genes. SNP data and cluster analysis indicated there was little population structure among the accessions, which were of highly diverse origin. Calculating nucleotide diversity at each site (π) and comparing the diversity of wild (*G. soja*) and elite (*G. max*) domesticated lines (selection index, or $\pi_{wild}/\pi_{domesticated}$) showed that many genes were candidate domestication loci. Classifying genes into three categories (duplicated, singleton, and orphan) revealed that orphan genes were overrepresented among genes putatively selected for in domestication. Among pairs of duplicated soybean genes, little correlation in selection index was observed among duplicate genes, indicating that domestication sweeps acted upon single genes and not families of genes. Finally, using short read mapping coverage to classify genes into 'core' and 'dispensable' gene sets showed that dispensable genes were more likely to have been through a domestication sweep, and that duplicated genes were less likely while orphan genes were more likely to have been selected in domestication. In all,

these results indicate that soybean's large genome is highly duplicated and that a small portion of these are dispensable genes, but that domestication sweeps likely acted disproportionately upon orphan genes and dispensable genes prone to presence-absence variation. These results underscore the role of dispensable genes and orphan genes, which likely arise from highly diverged ancient paralogs, in the domestication of soybean.

Materials and Methods

First, the most recent reference genome assembly and annotation of soybean (*Glycine max,* Wm82 v2 and Wm82.a2.v1) were obtained from Phytozome 12. A VCF (variant call file) file containing pre-computed SNPs and small indels for all 481 soybean lines (including 45 *Glycine soja* lines) in this study was obtained from Soybase (https://soybase.org/projects/SoyBase.B2014.02.php). With this variant data, the nucleotide diversity rate among the wild (*G. soja)* and cultivated (*G. max*) soybean lines was calculated by subsetting the VCF file into wild and elite datasets, and using the "sites-pi" function in VCFtools v0.1.16 (Danecek et al., 2011). The diversity was calculated across the length of the entire sequence of the gene using the coordinates from the GFF file marked "gene" from Phytozome 12. A 'selection index' for each gene was estimated by dividing the wild line diversity by the elite line diversity at that locus ($\pi_{wild}/\pi_{domesticated}$), such that genes with higher values for this number were putatively more strongly selected.

Soybean paralogs were extracted from gene families built in chapter 4 of this thesis. For each gene family containing multiple soybean genes, each paralog set was listed as a set of pairs, such that in a family with e.g. 3 soybean paralogs there were 3 pairings considered (1-2, 1-3, 2-3). A scatterplot of the values for the selection index for each gene pair, where the X coordinate was the value for one gene and the Y coordinate the value for its paired paralog, was created and

used to calculate a Pearson's correlation coefficient (r) for all paralog pairs. Genes were classified as either putatively selected or not, based on the distribution of selection indices for all genes. If a gene was in the top 5% of selection index values, it was putatively selected in domestication. All soybean genes were also classified into different duplication categories depending on their paralogs or orthologs within orthogroups. If a soybean gene had paralogs, it was considered to be "duplicated"; if a gene had orthologs but no paralogs, it was considered to be "singleton"; and lastly if a gene had no orthologs or paralogs in the species considered for the orthogroups, it was considered to be an "orphan".

To determine if presence-absence variation or dispensability also affected whether genes were selected in domestication, a method to measure absence via resequencing alignment to the Williams 82 reference was employed. A custom pipeline was built that, for each of the 481 soybean accessions considered here, downloaded the sequence reads from NCBI SRA (SRP06225 and SRP105183), aligned the reads to the reference using BWA-MEM with default parameters (Li and Durbin, 2009), and used the BEDtools (v2.28.0) (Quinlan and Hall, 2010) 'coverage' command to calculate the percent of nucleotides in each gene's coding sequence covered by at least one mapped read. A gene was considered "absent" if it was missing at least 50% coverage in the coding sequence in any accession. A gene was considered 'dispensable' if it lacked this same percent of coverage in any one accession among the data set. 1 accession's data had to be discarded due to zero reads mapping correctly to Williams 82, leaving a dataset of 480 accessions for this portion. Importantly, this method can only measure absence of a gene in a line compared to Williams 82, and cannot detect new genes not present in Williams 82.

Results

Diversity among 480 soybean and Glycine soja *lines*

In order to interrogate how domestication has affected the soybean genome, resequencing data for a wide variety of domesticated and wild soybean lines were obtained. In total, 480 lines were considered for this study from Soybase.org. The total dataset consisted of 481 lines, but one line was found to have erroneous or poor-quality sequence data, and was excluded from further analysis (PI518668, "TN 4-86"). 45 of these lines were *G. soja* accessions, and thus represented the wild and undomesticated or unbred germplasm of soybean. While the average coverage of the lines was generally 15x or 17x, 46 lines with diverse ancestry and pedigree were sequenced to a depth of 40x. Genotyping tables in the form of variant call format (VCF) files were obtained directly from the public data, which described single nucleotide polymorphisms (SNPs) called from short read alignments to the soybean reference genome of Williams 82 (Wm82 hereafter). This resulted in a total of 6,721,398 SNPs called among all 480 lines, with 0.53% mean heterozygosity (and 1.66% mean heterozygosity among *G. soja* lines). A principal components analysis revealed little population structure among the *G. max* lines, with only the *G. soja* wild lines clustering together when considered against the *G. max* lines (Figs 5.1,5.S2) – though some *G. max* lines did cluster together with the *G. soja* cluster through k-means clustering (k=3). Grouping accessions by their maturity group also showed little in the way of identifiable population structure (Fig 5.2).

Assessing domestication sweeps in duplicated and non-duplicated genes in soybean

The soybean genome is highly duplicated, with most genes having a sister paralog elsewhere in the soybean genome. Furthermore, these duplicates in soybean often maintain similar expression patterns in the same tissue type (Chapter 2), suggesting that they may also

maintain similar functions. Given these observations, it is possible that duplicate genes may have been selected at similar rates during domestication – that is, that two gene copies A1 and A2 were both selected for or subject to a bottleneck during domestication at the same time owing to their contribution to a common trait. To test whether duplicate pairs of soybean genes were selected through domestication simultaneously, sets of soybean gene paralogs to each other were identified via an orthogroups analysis using orthofinder (data from Chapter 4). These orthogroups identified paralog and ortholog relationships of all soybean genes and several other legumes, along with grape, resulting in a large set of gene families or orthogroups, allowing for an assessment of what genes had duplicate copies and what genes did not.

Using the SNP data from Soybase and the gene coordinates from the Wm82 soybean genome annotation version 2, the average nucleotide diversity (π) per gene was calculated with VCFtools v0.1.16 (Danecek et al., 2011). The sequences for the *G. soja* wild lines were separated from the domesticated *G. max* lines, and the average π for each gene from each wild or domesticated pool were calculated separately. A 'selection index' for each gene was estimated by dividing the π from the wild accessions by the π from the domesticated accessions ($\pi_{wild}/\pi_{domesticated}$) (Wang et al., 2017b) (Fig 5.3). This index served as a measure for the strength of selection on a gene during the domestication bottleneck. It would be expected that genes or loci in the genome that experienced strong selection during domestication would have greatly reduced genetic diversity in domesticated lines but much greater diversity among wild lines. Thus, a locus with a high $\pi_{wild}/\pi_{domesticated}$ would be a putative domestication site, having been subjected to a strong bottleneck. A pairwise examination of the selection indices across the chromosomes of the soybean genome indicated that there was little pattern to which chromosomes were selected more strongly than others, with only chromosome 18 showing any

evidence of stronger domestication selection than the others (Fig 5.S1). Putative domestication target genes here were defined as genes in the top 5% in their $\pi_{wild}/\pi_{domesticated}$ selection index values (Fig 5.3). The selection index was relatively closely distributed across chromosomes, but there were some differences detectable between the 20 chromosomes of soybean (Figs 5.5, 5.S1). Among the genes in in the soybean genome, one gene in particular, Glyma.10G090900.1, was an extreme outlier in this index with a value of $\pi_w/\pi_d = 1836.73$, while all other genes had $\pi_w/\pi_d < 500$. The closest *Arabidopsis* ortholog to this gene is a PIF1 helicase, which functions to assist in the maintenance and replication of nuclear and mitochondrial DNA (Byrd and Raney, 2017).

Next, a GO (gene ontology) term enrichment analysis was performed on the list of 2,552 genes identified as the top 5% of genes selected for during domestication using the Soybase GO enrichment tool (Morales et al., 2013). Among these putative domestication target genes, "defense response", "ADP binding", and "signal transduction" were found to be significantly overrepresented, with a Bonferroni-corrected p-value cutoff of 0.05 (Table 5.1). No other terms were significantly over- or under-represented among this set, indicating that either these putative domestication target genes were a relatively unbiased sample of the gene set in soybean, or that too few or too many genes were included in this list to give a significant result.

Using the list of duplicated genes derived from the Orthofinder (Emms and Kelly, 2015) analysis of sequenced *Faboideae* legumes, each gene in the soybean genome was classified as either "duplicated," "singleton," or "orphan". Duplicated genes were those with at least 2 total identifiable copies or paralogs in soybean. Singleton genes were genes with identifiable orthologs to at least one of peanut (*Arachis hypogaea*), diploid wild peanut (*Arachis duranensis* and *Arachis ipaensis*), *Medicago truncatula*, common bean (*Phaseolus vulgaris*), or the non-legume grapevine (*Vitis vinifera*), but no paralogs in soybean (i.e. single copy in soybean).

Orphan genes were genes with no identified paralogs in soybean nor orthologs to the other species. Notably, orphan genes are thought to often arise from highly diverged paralogs or ohnologues (previously duplicated genes), and thus may represent an extreme case of neofunctionalization, and are legacies of duplication (Tautz and Domazet-Lošo, 2011). In all, 38,320 genes were classified as "duplicated", 3,849 as "singleton", and 8,849 genes as "orphan". Comparing the set of the bottom 95% of domesticated genes to the top 5% identified earlier shows that there are fewer duplicated genes, fewer singletons, and more orphan genes proportionally among the top 5% of domestication target genes (Fig 5.4). A chi-squared test revealed that these differences were statistically significant, with orphan genes being overrepresented and the other two categories being underrepresented among the top 5% of selected genes ($p < 2.2e-16$, 2 d.o.f.). However, duplicated genes still dominated overall, as most genes in soybean are duplicated.

During the process of domestication, loci controlling a spectrum of traits that comprise the domestication syndrome were intentionally or inadvertently selected upon by humans. However, in nearly all major crops, domestication was undertaken solely via selection on phenotypes, as these events generally happened about 10,000 years ago, far before any knowledge of genetics had been developed (Smith and Nesbitt, 1995). It is possible, then, that in a case where several loci may control a single domestication trait, those loci could have been selected simultaneously. In the soybean genome, many genes have retained duplicates with similar expression patterns across or within tissue types (e.g. gene copy "A1" has the same or similar expression levels as gene "A2" in leaf tissue), suggesting that duplicate genes in soybean may retain similar functions (Roulin et al., 2013). This is particularly true of more duplicates arising from the most recent whole genome duplication (WGD) event in soybean, ~8-13 Mya.

Thus, it is possible that duplicate copies (ohnolog) of genes in soybean may have been selected simultaneously during domestication, owing to their high sequence and functional similarity. To investigate this possibility, pairs of duplicate genes in soybean were constructed from the legume and grape orthogroups. For every soybean gene with a paralog, every pairing of that gene with all its paralogs was listed. For example, a gene with copies A1, A2, and A3 was paired A1-A2, A1-A3, and A2-A3. Then, the selection index was compared between each of these pairs of paralogs for all genes in the soybean genome, and the results were plotted as a scatterplot to determine whether the selection index or strength of domestication for duplicated genes was correlated (Fig 5.5). The Pearson's correlation coefficient was $r = 0.0413$ ($p = 3.89e\text{-}44$), indicating that the selection index of a given gene bore little to no relation to that of its copies.

Identifying core and dispensable gene sets among soybean lines

SNPs are common throughout genomes and are often the primary variant used to distinguish individuals in a population or diversity panel in modern breeding programs, but they are only a subset of the possible variation in a genome. Small insertions and deletions (indels) are also common like SNPs, and are often included when genotyping accessions in modern breeding programs. However, both of these smaller-scale variations in nucleotide sequence together still fail to capture the breadth of diversity in an organism. In fact, there are many large sequences in varying genotypes among plants and other organisms that are not shared with their relatives (Li et al., 2014; Pinosio et al., 2016; Vernikos et al., 2015). These structural variants can be hundreds to millions of base pairs in size, and can involve large insertions, large deletions, inversions, tandem duplications, and more. In addition, these variants can include entirely new genes or gene copies not shared with a reference genotype. Large scale, deep sequencing with long read lengths or advanced mapping techniques can help to resolve these non-homologous

structural variants without the need to align to a reference genome (Schadt et al., 2010). However, these are currently very expensive and time-consuming, and are thus not feasible for many studies. Regardless, it is possible with short-read sequencing data to detect simple structural variants like the absence of a gene in a non-reference genotype. Duplicate genes often evolve rapidly and diverge greatly from their paralogs, or can even be deleted or pseudogenized, as indicated by the presence of a considerable amount of orphan genes in soybean. Thus, it is worth investigating how different genes may be present or absent (presence-absence variation, or PAV) in diverse soybean accessions, which would indicate ongoing variation in the gene set of soybean.

To accomplish this, all short-read sequencing data for the 480 soybean lines were downloaded from the NCBI short read archive (SRA) database. Then, each sequencing run was aligned against the Wm82 reference genome using BWA MEM v0.7.17-r1188 with default parameters (Li and Durbin, 2009). These alignments were then analyzed using the "bedtools coverage" tool (v2.27.1) and the gene coordinates from the soybean annotation version 2.0 in order to calculate the percent coverage for each gene in the Wm82 reference genome for each sequenced accession. The percent coverage was defined as the number of nucleotides within the defined gene coordinates that were covered by at least one sequence read from a particular accession. A given gene was considered "dispensable" if it had 50% coverage or less in at least one accession, and "core" if it had more than 50% coverage in all accessions sequenced.

This resulted in 2,792 (5.5%) dispensable genes (or 972 if "dispensability" was defined as having 0% coverage in at least one accession) and 48,226 (94.5%) core genes. Of the genes defined as dispensable, the vast majority were missing in only one line, while 18 genes were missing in every line but present in Williams 82 (Fig 5.6). A GO term enrichment analysis

indicated that genes with the GO annotations “defense response,” “signal transduction,” “systemic acquired resistance,” and “protein phosphorylation” were overrepresented among dispensable genes with a Bonferroni corrected p-value of < 0.05 (Table 5.2), terms commonly associated with resistance genes.

Disentangling gene dispensability and duplication, and domestication in soybean

Duplicate genes often diverge rapidly from their sister genes, and sometimes are lost altogether (Panchy et al., 2016). As stated earlier, orphan genes are perhaps one example of this process at work, where some genes appear to have no homology to any other genes within or between species and are likely highly diverged ancient duplicate genes. While divergence is one possible outcome for a duplicate gene copy (i.e. subfunctionalization or neofunctionalization), deletion or pseudogenization is another. The processes by which these genes diverge, and which determine their ultimate fate are thought to take place on long time scales, often millions of years. However, domestication imposes a severe population and genetic bottleneck on a plant genome, and could induce drastic changes in the genome in the space of only a few thousand years or even a few generations. By comparing how genes of different duplication status (duplicated, orphan, or singleton) and dispensability (core or dispensable) differ between elite and wild soybean lines, and how these genes may have been selected through domestication, a better picture of how domestication and improvement have been affected by human intervention can be discerned.

All genes in the soybean genome were classified as duplicated, singleton, or orphan; as dispensable or core; and as selected through domestication (top 5% of selection index) or not selected. In general, orphan genes were more likely to be dispensable than duplicated genes or singletons (Fig 5.7). A two-way ANOVA was performed on all genes which treated selection

index as a dependent variable while duplication status and dispensability were treated as independent variables. Both dispensability and duplicability were found to significantly affect selection index: dispensable genes were more likely to be selected in domestication, and orphan genes were more likely to be selected than duplicated or singleton genes (Fig 5.8a). Considering duplication or dispensability separately showed similar patterns as before: orphan genes and dispensable genes were disproportionately higher in their selection index than their duplicated, singleton, or core counterparts (Fig 5.8b-c). When the genes were also partitioned by their presence in the top 5% of selected genes, dispensable genes were once again shown to be overrepresented in the set of genes that were most strongly selected for during domestication (Fig 5.9a-b). Furthermore, orphan genes were overrepresented in the top 5% of selected genes and were overrepresented within the dispensable genes in the top 5% of selected genes (Fig 5.9c). Overall, these results show that orphan genes and dispensable genes were more strongly selected in domestication and are overrepresented among the genes most strongly selected for in domestication.

Discussion

Duplication, whether at the whole-genome, segmental, single gene, or short repeat scale, has long been known to be an important evolutionary force on plant genomes. Comparatively less is known, however, how these duplication processes, or results thereof, may be at play among elite breeding lines and the wild relatives of crop plants. Duplication can serve to modify genes, expand gene families, create new functions, or create entirely new genes from preexisting genome segments. The evolutionary versatility of duplication is widely appreciated (REFS?), but its importance in breeding, domestication, and improvement of crop plants has yet to be substantially described or exploited. This work aimed to examine whether duplicate gene copies

were selected at similar rates during domestication in the highly duplicated soybean genome, and whether duplicate genes contribute to the varying gene complement of different soybean accessions.

Soybean domestication: the bottleneck from G. soja to G. max

Soybean was probably first domesticated about 3,500 years ago in East Asia, most likely near the Yellow river basin (Sedivy et al., 2017). It is thought that ancient humans created what is now known as modern cultivated soybean, *Glycine max*, from a native wild progenitor, *Glycine soja*. In contrast to the domesticated soybeans farmers and consumers know today, *G. soja* has a vine-like twining growth habit, small and hard black seeds, and is smaller overall (Kuroda et al., 2013; Sedivy et al., 2017). Despite its agronomic shortcomings, the wild *G. soja* is an important source of genetic diversity for soybean. Soybean has a particularly narrow genetic base as compared to other important crops, and this is most pronounced within the North American soybean germplasm, where just 80 ancestral lines account for an estimated 99% of the parentage of North American soybean accessions (Gizlice et al., 1994). Thus, the domestication bottleneck for soybean was especially strong when compared to other crops, and this is borne out in our results: no particular regions or genes in the soybean genome (Fig 5.3), and few to no particular GO terms, dominated the highly selected genes (Table 5.1). This suggests that the domestication bottleneck in *G. max* may have acted strongly across essentially the entire genome and gene set, leaving very little diversity at any locus as compared to wild lines. This highlights a daunting problem for soybean breeders as genetic diversity is the greatest source of any new or improved traits for any crop. Introgressing valuable traits like cold hardiness or disease resistance from wild *G. soja* or *G. max* landraces is possible, but crossability between *G. max* and *G. soja* is variable, and there may be linkage drag between undesirable traits and desirable

traits from the wild parent (Singh and Hymowitz, 1988). Understanding the breadth of genetic diversity that currently exists in both the wild and domesticated soybean germplasm is therefore critical for employing this germplasm in the continued improvement of the crop; what variation there already is in elite soybean lines becomes far more valuable in light of such a narrow genetic base. Until recently, most genetics and breeding studies in plants have focused heavily on small variants like SNPs or microsatellite repeats, but newer sequencing technologies and techniques have allowed for the resolution of larger variants which might comprise entirely new genes or gene families not present in other genotypes in a crop germplasm. These structural variants, especially presence-absence variants, can be valuable and untapped sources of genetic diversity for crop improvement.

Gene and genome duplication are a vital source of genetic diversity and even speciation, not only for plants but other kingdoms (Ohno, 1970). Entirely new genes or genes with new or modified functions can arise from duplicated gene pairs, since often one gene copy experiences relaxed selection and is thus free to mutate and generate new functions or modified functions, since its 'normal' sister copy can still fulfill the original function. Sometimes, however, these new functions are deleterious, or the new copy is not sufficiently beneficial to the plant, and these genes are either deleted or accumulate enough mutations to become nonfunctionalized or pseudogenized. Soybean has at least three whole-genome duplication events ~130 My, ~60 My, and ~13 My ago along with ongoing 'background' events like tandem duplications and transposon movement (Shoemaker et al., 2006). Gene duplicates may be millions of years old, or very recent, and may have played roles in the domestication of soybean in the past few thousand years. Studying the germplasm of soybean, both wild and domesticated, can illuminate not only

the timescale at which duplicate gene evolution happens, but also provide insights as to how these duplicate genes may have played roles in the domestication of soybean.

In this study, the effects of domestication were assessed through a simple measure of the reduction in diversity at a given gene between wild *G. soja* lines and domesticated *G. max* lines ($\pi_{wild}/\pi_{domesticated}$). To create a simple cutoff, the genes within the top 5% of values for this 'selection index' were considered to be putative domestication targets, as domestication syndrome targets a wide range of traits that may be quantitative or polygenic in nature. Genes with functions of "defense response", "ADP binding", and "signal transduction" were overrepresented among these putative domestication target genes, with no other terms being significantly enriched or depleted. That so few GO terms were significantly enriched or depleted among domestication targets could indicate that domestication acted largely randomly upon the soybean genome, that domestication-related genes have highly diverse functions, or that gene ontology terms do not necessarily predict the kinds of functions that genes involved in domestication. Perhaps, then, changing the cutoff for domestication targets in a large genome like soybean's to the top 1% of selection indices may give different results, at the expense of potentially excluding many genes which were indeed important in domestication since, as discussed in more detail below, other studies have estimated up to 4% of loci are involved in domestication (Wright et al., 2005). In addition, the one gene with by far the highest selection index (Glyma.10G090900.1) appeared to be involved in DNA maintenance and replication, which is not commonly associated with domestication traits like shattering or upright growth habit. This method, then, may not be useful in predicting specific genes that control domestication syndrome traits. However, this method did offer other insights into the broad types of genes in the soybean genome that were selected on for thousands of years, especially

when considering a larger number of loci than the scant few that are often hypothesized to be sufficient for domestication.

It is important to note that the method of defining 'selection during domestication' or a bottleneck as (π_w/π_d) suffers from a basic mathematical problem: if a given locus or gene has 0 nucleotide diversity in the domesticated lines, the selection index cannot be calculated as this would necessitate division by zero. In the data here, no genes in soybean had zero diversity among elite lines, but 6,544 genes had $\pi = 0$ in wild lines – hence why the selection index was not the inverse, π_d/π_w. This may be due to poor sequence quality in these lines, importantly, and may not represent the true diversity of wild soybean. Furthermore, the prevalence of genes with $\pi = 0$ in wild lines could also be a result of there being only 45 *G. soja* lines in the dataset, while domesticated *G. max* lines had 435 lines, meaning there were more opportunities for *de novo* mutations to have arisen in these lines in any given gene and thus less chance that a gene had zero nucleotide diversity.

The importance of orphan genes: relics of duplications past

One of the types of genes overrepresented among domestication targets was orphan genes. Orphan genes are a unique case in studies of homology between or within genomes, as they do not have any homologs in their own species or to any other species. It has been proposed before that orphan genes commonly arise out of duplication events and may be extreme examples of neofunctionalization. In this scenario, a gene "A" could copy into duplicates "A1" and "A2". A1 might maintain the ancestral function of "A", leaving A2 free to 'explore' evolutionary functions, gaining new functions, modifying old functions, losing functions, or being lost altogether. In this process of relaxed selection and accumulation of new mutations, the duplicate may change so much relative to its original sister gene that it is no longer recognizable

as a homolog (Bellora et al., 2008; Domazet-Loso and Tautz, 2003). This would generate an orphan gene despite that gene having once been a copy of another gene in the same genome. Thus, orphan genes may actually represent an important example of duplicate gene evolution. It is worth noting, however, that orphan genes can also potentially arise *de novo* from non-coding genome sequence, whether transposons or simple repeats – though it is estimated that only 5% to 10% of orphan genes arise this way. Furthermore, some orphans may just be artifacts of poorly-sequenced lineages or otherwise incomplete data, or could even simply be transposon-related sequences misannotated as genes (Bellora et al., 2008; Donoghue et al., 2011). Future work should include BLASTing these orphan genes against all species in e.g. GenBank, searching these genes against RepBase or other transposon database, or comparing them to a known list of transposon-related genes to determine the validity of the orphan gene set. Previous work has found that the number of novel TE insertions in diverse *G. max* soybean accessions is about 300 per accession, and that most of these are in non-genic regions, which would suggest that these probably do not account for all of the 8,852 orphan genes found here (Tian et al., 2012).

Here, orphan genes were found to be overrepresented among genes showing evidence of selection during domestication (i.e. genes in the top 5% of π_w/π_d) and were also overrepresented among dispensable genes (Fig 9). This means that genes which may have begun as duplicates millions of years ago could have evolved to become genes that played critical roles in the domestication of soybean. Originally, it was thought that changes in just a few genes with large effect were sufficient to achieve domestication in crops, as evidenced in e.g. rice (Cai and Morishima, 2002), wheat (Peng et al., 2003), tomato (Khan et al., 2019), and maize (Buckler et al., 2001) . However, newer genomics-enabled studies have indicated that perhaps larger portions of the genome are implicated in domestication, and domestication may be a result of both a few

large-effect loci and many small-effect loci acting in tandem to produce a domesticated plant (Purugganan and Fuller, 2009; Sedivy et al., 2017). In fact, 2-4% or more of the loci in a genome could be putatively involved in domestication (Wright et al., 2005). Thus, the relatively large number of genes categorized here as most strongly selected during domestication (5% or about 2500 genes), and the large amount of variably-present orphan genes therein, could still yield potentially important genes as targets of domestication and perhaps further improvement. Known, cloned genes associated with domestication in soybean like *E1*, a maturity locus, and its homologues are variable in their duplication status: for *E1La* and *b*, both known paralogs (Glyma.04G143300.1 and Glyma.04G156400.1) are contained in one gene family and would thus be of the "duplicated" class here (Xu et al., 2015). A thorough meta-review of the literature of cloned domestication genes and their duplication status would be revealing in assessing the association between orphan genes and domestication sweeps found in this study.

It is also possible, however, that these orphans had lower diversity in elite lines than duplicate genes due to their propensity for dispensability (Fig. 5.7). This would mean that the few wild lines that gave rise to the elite soybean germplasm might have contained a dispensable orphan gene, whereas the other wild lines may have been missing the gene altogether. Thus, recombination events over successive generations could not have introduced variability back into these genes. This would lead the orphan genes to appear to have been selected strongly during domestication, when they simply were selected from a smaller starting gene pool among wild soybean progenitors than their duplicate gene counterparts, which were more well-retained (Chapters 2 and 3).

The core and dispensable gene sets in soybean, and their relationship to duplication and domestication

Reference-based SNP and small indel calling approaches to defining diversity in a species have an inherent drawback in that they generally cannot detect large DNA sequences either inserted or deleted. Thus, different approaches are needed to fully describe the variation within a species' germplasm. Since long-read technologies are prohibitively expensive for resequencing a large panel like the 480 lines in this study, a simpler coverage-based approach employing more accessible short-read data was used to determine whether genes were core or dispensable. Genes with less than 50% coverage in any line were defined as 'dispensable', a similar approach to other studies which use e.g. reciprocal mapping from a reference to a non-reference genotype and a 75%/25% reciprocal mapping threshold (Hirsch et al., 2016). Even with the somewhat more sensitive absence-only approach used here, only 5.5% of genes were defined as 'dispensable' of the over 55,000 genes in the soybean genome. This is similar to core and dispensable gene sets found in more in-depth cultivar comparisons in other species, such as maize where 2,713 (about 7%) of the total genes were found to be unique to either PH207 or the reference B73 genotypes (Hirsch et al., 2016). This maize comparison, however, was accomplished via intensive assembly and annotation of a single non-reference accession. Thus, while the approach here is incomplete, and a more thorough approach would involve *de novo* assembly and gene annotation for every soybean line considered here, the simple short-read mapping absence approach still defined a core and dispensable gene set in line with what has been seen in other studies. Additionally, assembling and annotating just 7 *Glycine soja* lines in a previous study yielded an estimated 80% core and 20% dispensable gene set for diverse wild

soybean lines, indicating that the core and dispensable genes defined in this study may still be a small subset of the total variable genic content between soybean lines (Li et al., 2014).

The dispensable soybean genes were found to be overrepresented among genes showing evidence of having been bottlenecked during domestication. Furthermore, these dispensable genes were also overrepresented among orphan genes, which were also more likely to have experience a domestication bottleneck (Figs 7 & 9). This suggests that orphan genes, which are possibly highly diverged relics of duplication events, may be important drivers of the agronomic traits humans selected for in soybean over thousands of years. Additionally, these orphan genes may be present or absent among different elite soybean lines, indicating that there is still considerable variation in orphan and/or dispensable genes that may affect valuable agronomic traits, and that there still may be more improvement breeders can obtain by combining lines with or without genes that affect these domestication target traits, utilizing complementation of gene absence. The results here are not conclusive on this matter, however, and a closer examination of the individual orphan, dispensable, and bottlenecked genes is needed to determine which genes truly are targets for improvement and which genes may simply be genetic hitchhikers of the 'real' domestication targets. Furthermore, knowledge of how the presence or absence of a given candidate gene affects the phenotype of the plant is needed before breeding efforts can be undertaken, which again necessitates even more investment into understanding these genes before breeders can make much-needed gains in soybean. The most important question to answer on this topic is whether traditional SNPs or satellite markers can be strongly linked to structural variants, and whether that means that extensive resequencing efforts are warranted in the first place.

For a crop like soybean that is highly inbred, with long haplotype blocks in full linkage disequilibrium (LD) of sometimes 200kbp or more (Contreras-Soto et al., 2017), it is reasonable to assume that in many cases, structural variants like PAV or CNV will be linked strongly to SNPs which are detectable with traditional methods. This would, on its face, indicate that these structural variants are not worth considering in a breeding program. However, the results of this study and other studies in e.g. humans indicate that structural variants like PAV are often rare variants – the kind that drive important QTL of great interest to breeders, and which are often missed in SNP-based association studies e.g. GWAS (Bernardo, 2016) . Instructively, even in this study, over 2000 genes were missing in just one of the 480 accessions analyzed (Fig 5.6). Furthermore, important structural variants may not always be in strong LD with any nearby detected SNPs (Hinds et al., 2006). Even accounting for strong LD in soybean, an important source of variation is perhaps being missed, and this structural variation could even have greater phenotypic impacts than e.g. SNPs (Torkamaneh et al., 2018). However, before this can be acted upon, the linkage between SNPs or small indels and structural variants like PAV must be determined. The data in this study already allow for this kind of analysis, and it would be an important avenue of future investigation into this question. If the linkage between the PAV variants described here and common SNPs is found to be weak in some cases, future soybean breeders might do well to take heed of the important structural variation that may be missed with SNPs or microsatellites. Assessing how often structural variants in soybean are linked tightly with known SNPs or QTL-associated SNPs would thus be essential for future work in answering this question.

Conclusions

Land plants have thrived on Earth for almost a billion years (Knauth and Kennedy, 2009). In this time, plants developed widely diverse structures, life cycles, and behaviors. While natural selection has created a staggering assortment of plant species and genotypes, human-imposed artificial selection via domestication accelerates the process of evolution immensely; some plants like *Lupnius luteus* have even morphed from wild, intractable forms to non-shattering, nutritious cultivated forms in just 16 years (Hanelt, 1986). Soybean was domesticated just a few thousand years ago, a mere moment on an evolutionary timescale. In this time, however, enterprising breeders and farmers have managed to tame a vine-like, delicate plant with small, hard black seeds into a robust, easy to manage plant and an economic powerhouse providing a major source of protein, oil, and animal feed for the world. This process, however, comes with drawbacks, and in soybean's case, it was a considerable genetic bottleneck. Modern elite soybean lines in some gene pools can trace their ancestry back to just a few dozen parents, while wild soybean lines show surprising variation in their morphology, development, and resistances to biotic and abiotic stress. Any further improvements to the agronomy of soybean thus require not only broadening the genetic base of elite soybean germplasm, but for a deeper understanding of the architecture of genetic and phenotypic diversity in soybean. While it has been well-known that soybean's genome is highly duplicated, little is known about how the duplicated nature of soybean has affected its domestication and improvement as a crop. Furthermore, the progress of sequencing technologies has allowed for cheaper, more accessible sequencing and resequencing of soybean, and thus has enabled approaches to assessing diversity outside of the strict confines of e.g. SNP or microsatellite markers, even enabling the identification of entirely new genes, or genes that are missing in one or many accessions (presence-absence variation). This study's results suggest that orphan genes, with no homologs to any other genes within or between

species, and which are commonly thought to arise as highly diverged relics of ancient gene duplications, are perhaps critical to domestication traits. Furthermore, these genes were also commonly variable in their presence or absence in different soybean lines, indicating that the evolution of these orphan duplicates was variable and likely affected important domestication traits in soybean. Overall, these results indicate that soybean's duplicated nature and the genetic variation it induced may have been crucial to its success as a crop worldwide.

Table 1. GO terms most significantly enriched (Bonferroni corrected p-value < 0.05) among genes in the top 5% of selection index (π_w/π_d). Terms not below the significance cutoff are not included.

GO ID	Description	Status	Corrected P-value
GO:0006952	defense response	Overrepresented	8.16E-17
GO:0043531	ADP binding	Overrepresented	8.41E-11
GO:0007165	signal transduction	Overrepresented	1.65E-09
GO:0006355	regulation of transcription, DNA-dependent	Overrepresented	0.887350609

Table 2. GO terms most significantly enriched (Bonferroni corrected p-value < 0.05) among dispensable genes (genes with at least one accession with 50% or less coverage).

GO_id	Description	Status	Corrected p-value
GO:0006952	defense response	Overrepresented	1.93E-58
GO:0007165	signal transduction	Overrepresented	5.01E-26
GO:0009627	systemic acquired resistance	Overrepresented	0.00700815
GO:0006468	protein phosphorylation	Overrepresented	0.0081004
GO:0048544	recognition of pollen	Overrepresented	0.01770424
GO:0006995	cellular response to nitrogen starvation	Overrepresented	0.01962464
GO:0006865	amino acid transport	Overrepresented	0.04751686

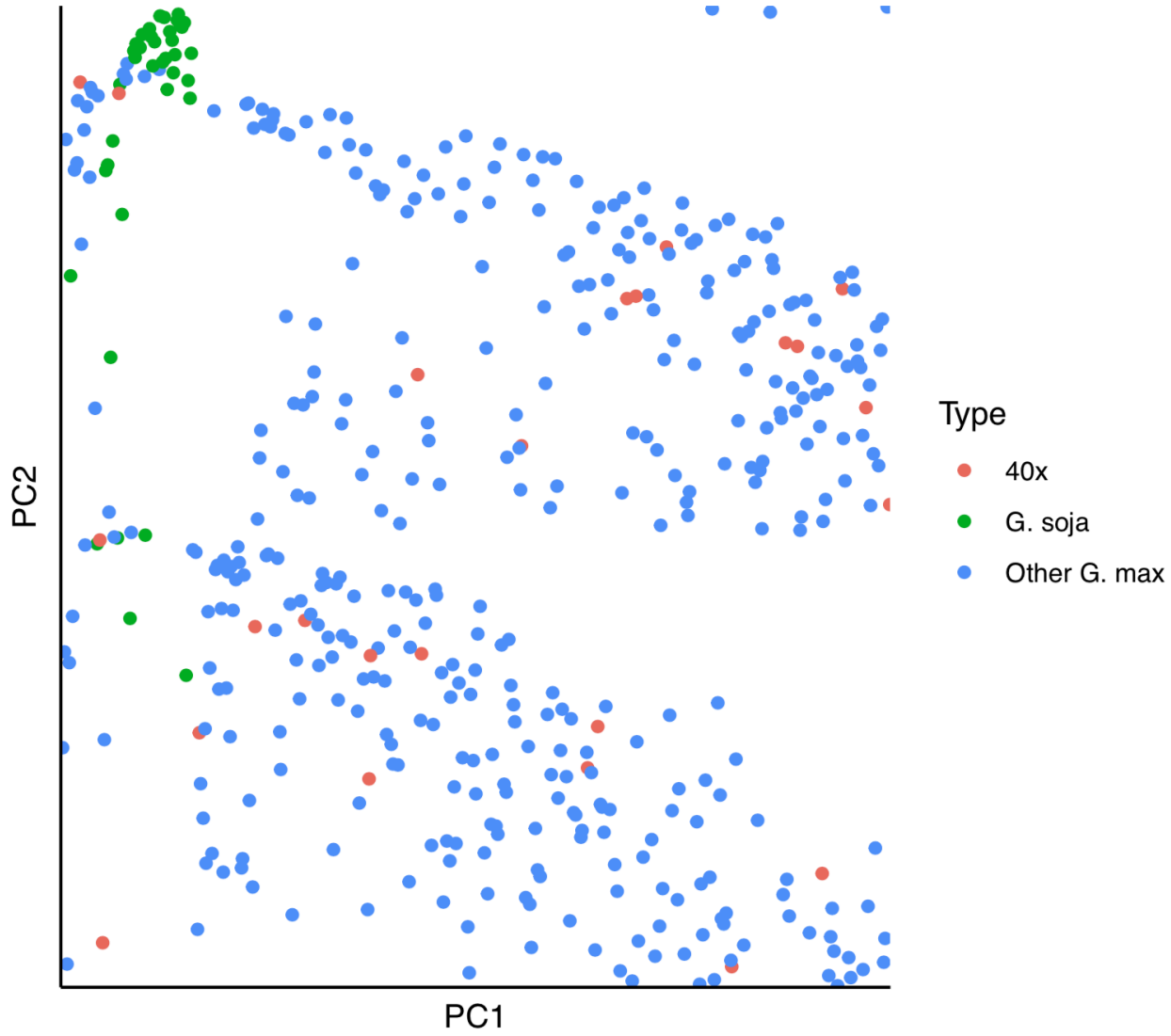

Figure 5.1. Scatterplot of the first two principal components of the 480 soybean accessions, grouped by their coverage or species.

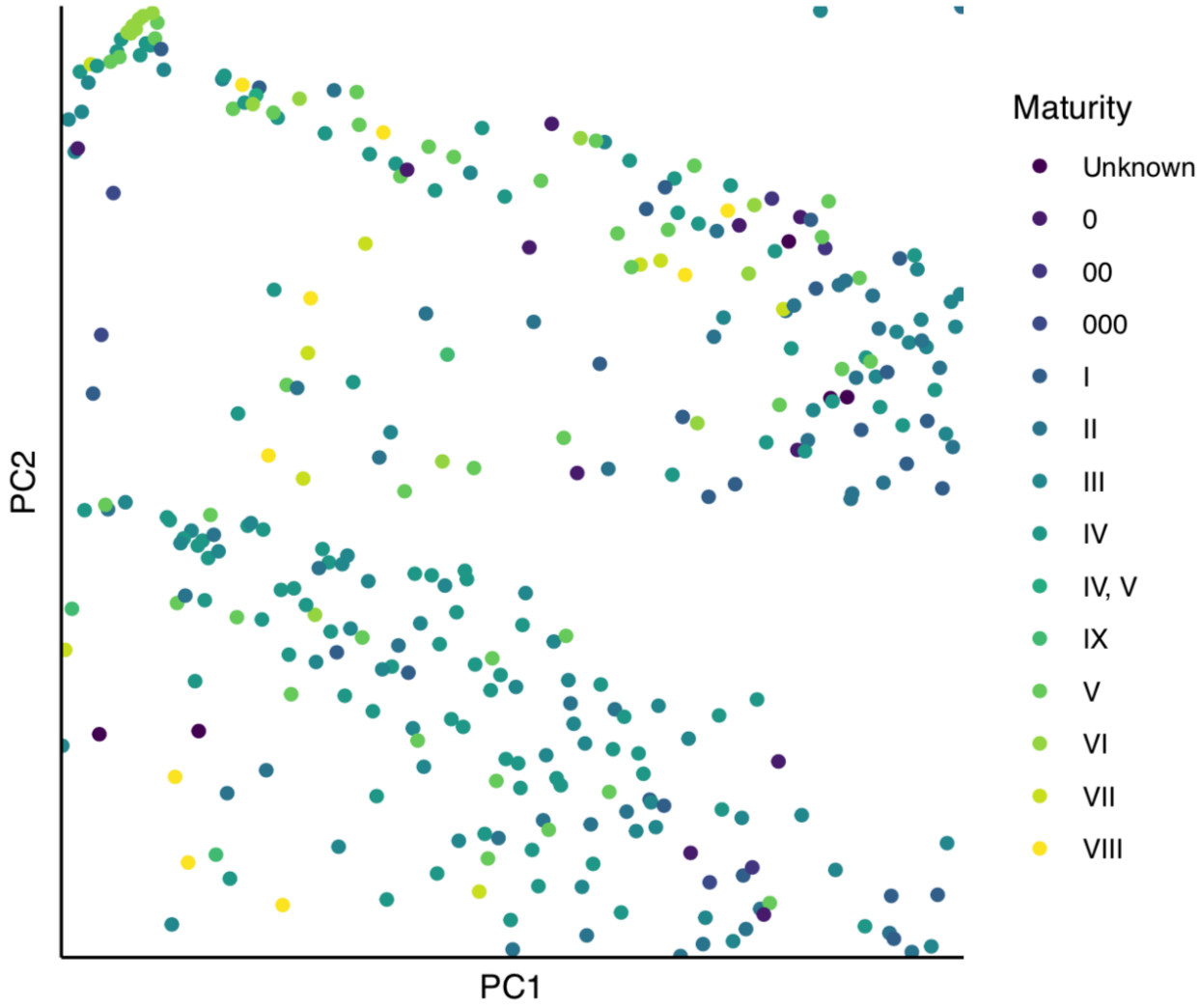

Figure 5.2. Scatterplot of first 2 principal components of the 480 soybean accessions grouped by their maturity group. The higher numbers indicate accessions with later maturity.

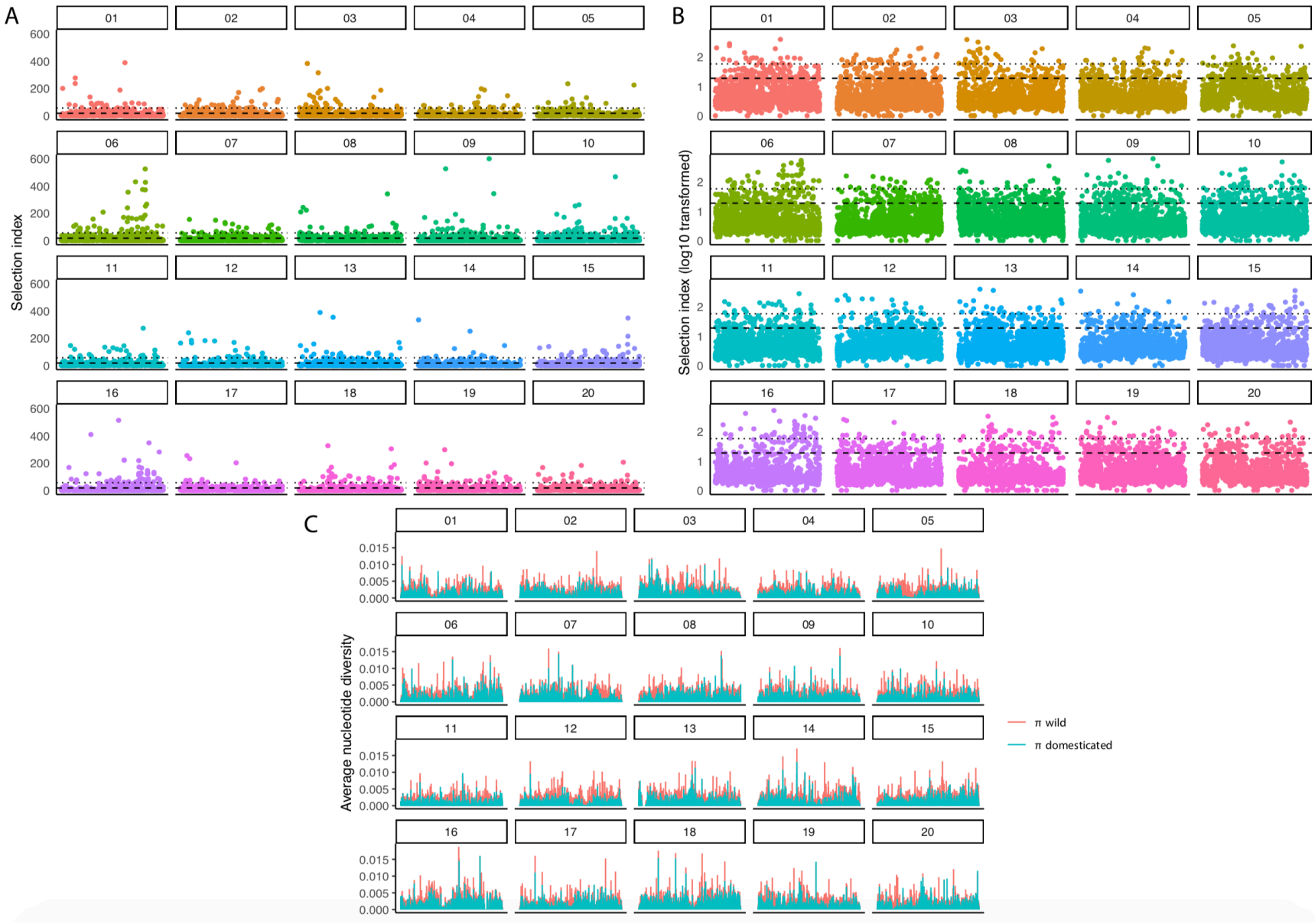
A
B
C
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
Selection index
Selection index (log10 transformed)
Average nucleotide diversity
π wild
π domesticated

Figure 5.3. Distribution of selection indices and nucleotide diversity of genes among the 480 accessions across the 20 soybean chromosomes. A) Selection index (π_w/π_d) for all genes across all 20 chromosomes. The lower dashed line is the cutoff of the top 5% of values, and the dotted, top line is the cutoff for the top 1% of values. B) Log-transformed selection index (π_w/π_d) (same data as panel A). Dashed and dotted lines represent the top 5% and top 1% of values cutoffs respectively. C) Average wild and domesticated nucleotide diversities (π) at each gene.

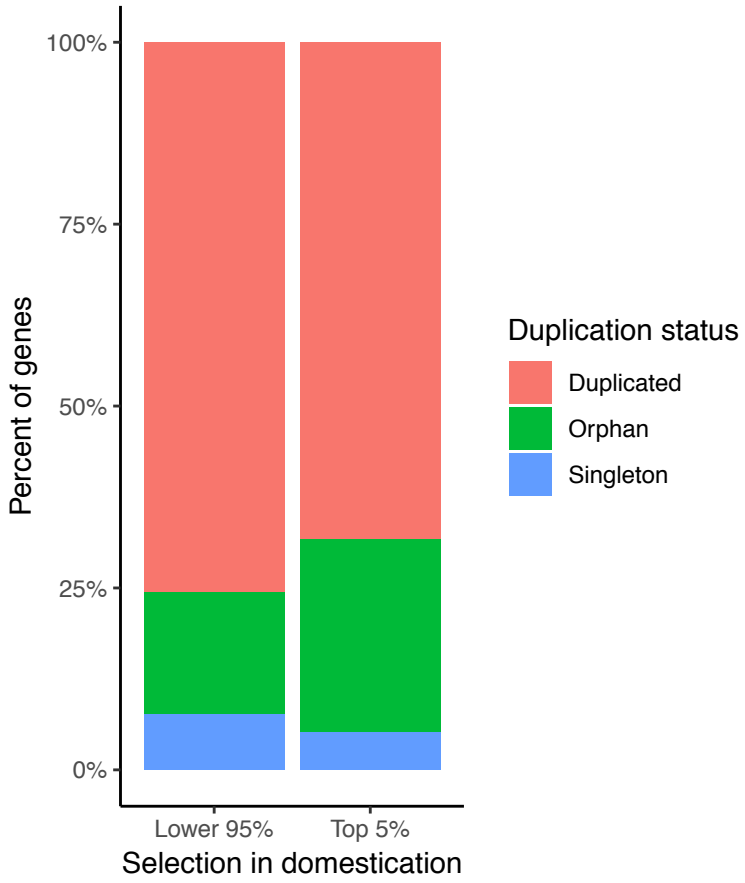

Figure 5.4. Percent of genes belonging to each duplication category among genes with the strongest evidence of selection in domestication (top 5% π_w/π_d) and non-selected genes.

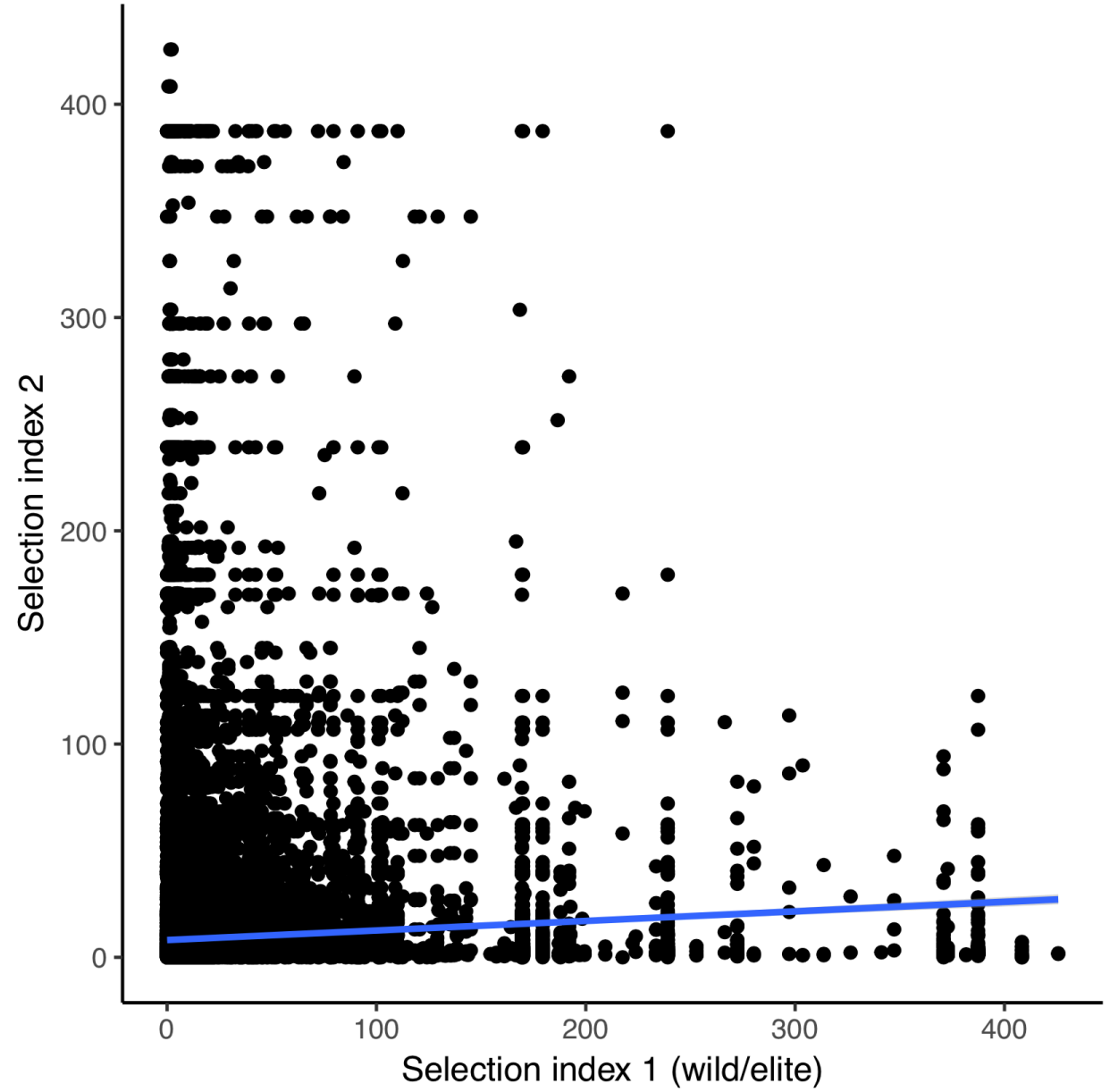

Figure 5.5. Scatterplot of selection index values for pairs of paralogous genes in soybean. The blue line represents the least-squares regression fit line (r = 0.0413). Genes related to Glyma.10G090900.1, which had a selection index of over 1800, are excluded to preserve scale.

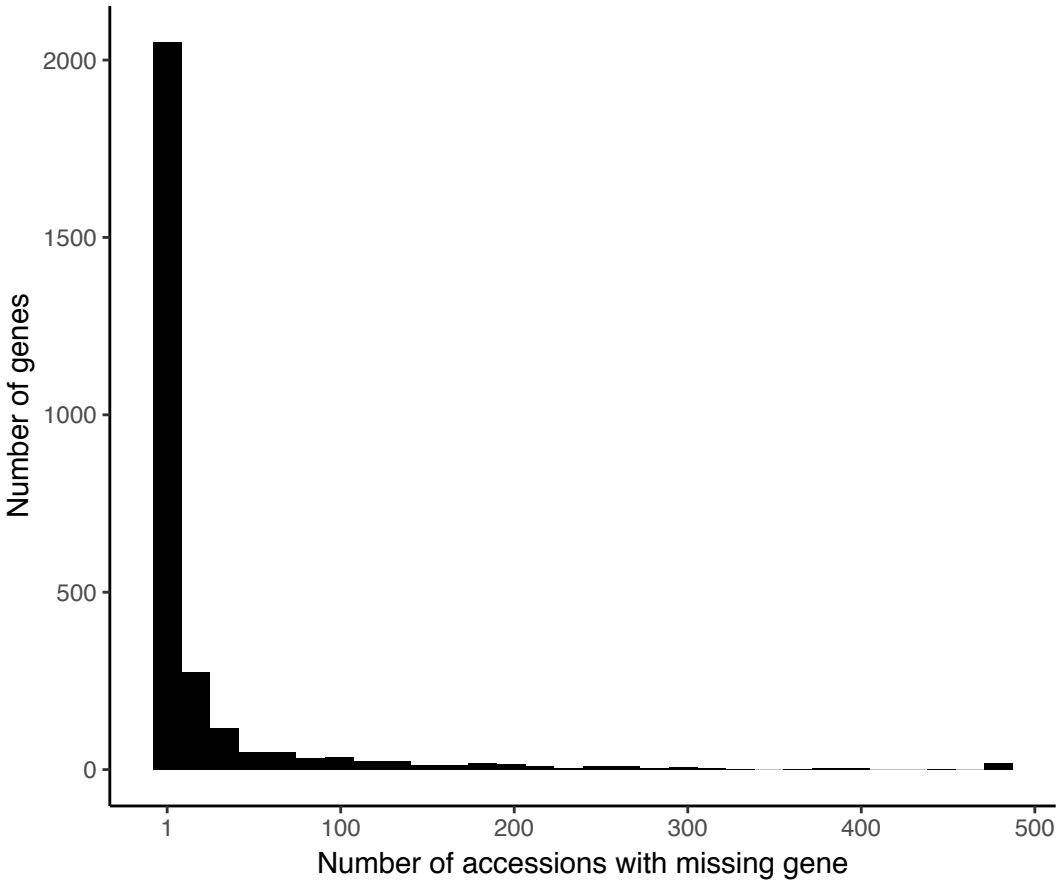

Figure 5.6. Histogram of the number of accessions in which every dispensable gene is missing.

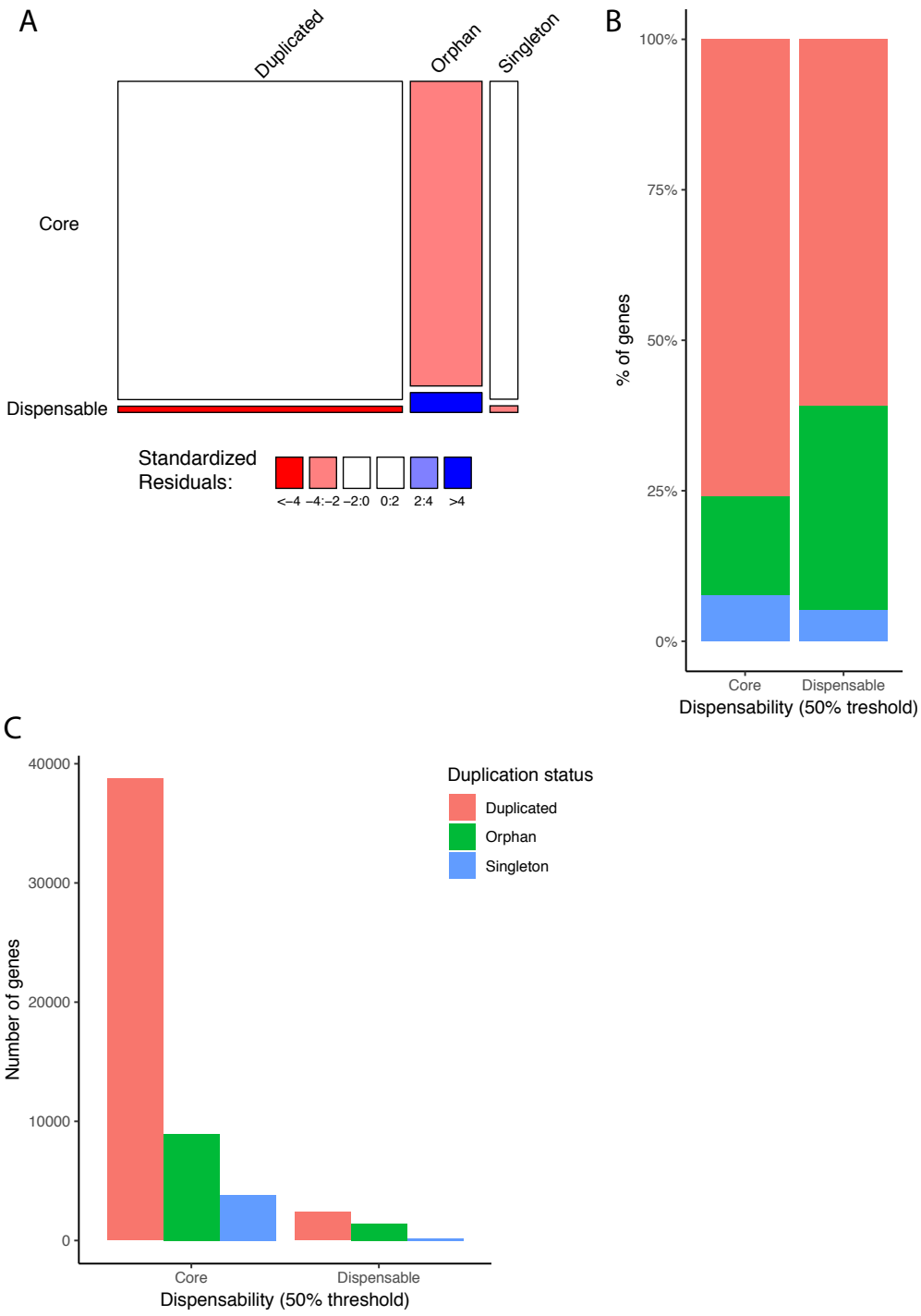

Figure 5.7. Relationship of dispensability and duplication status. A) Mosaic plot of dispensable and duplicated genes. Boxes on the bottom indicate residuals (over- or under-representation). B) percentage and C) raw counts of genes in each core/dispensable and duplicated status.

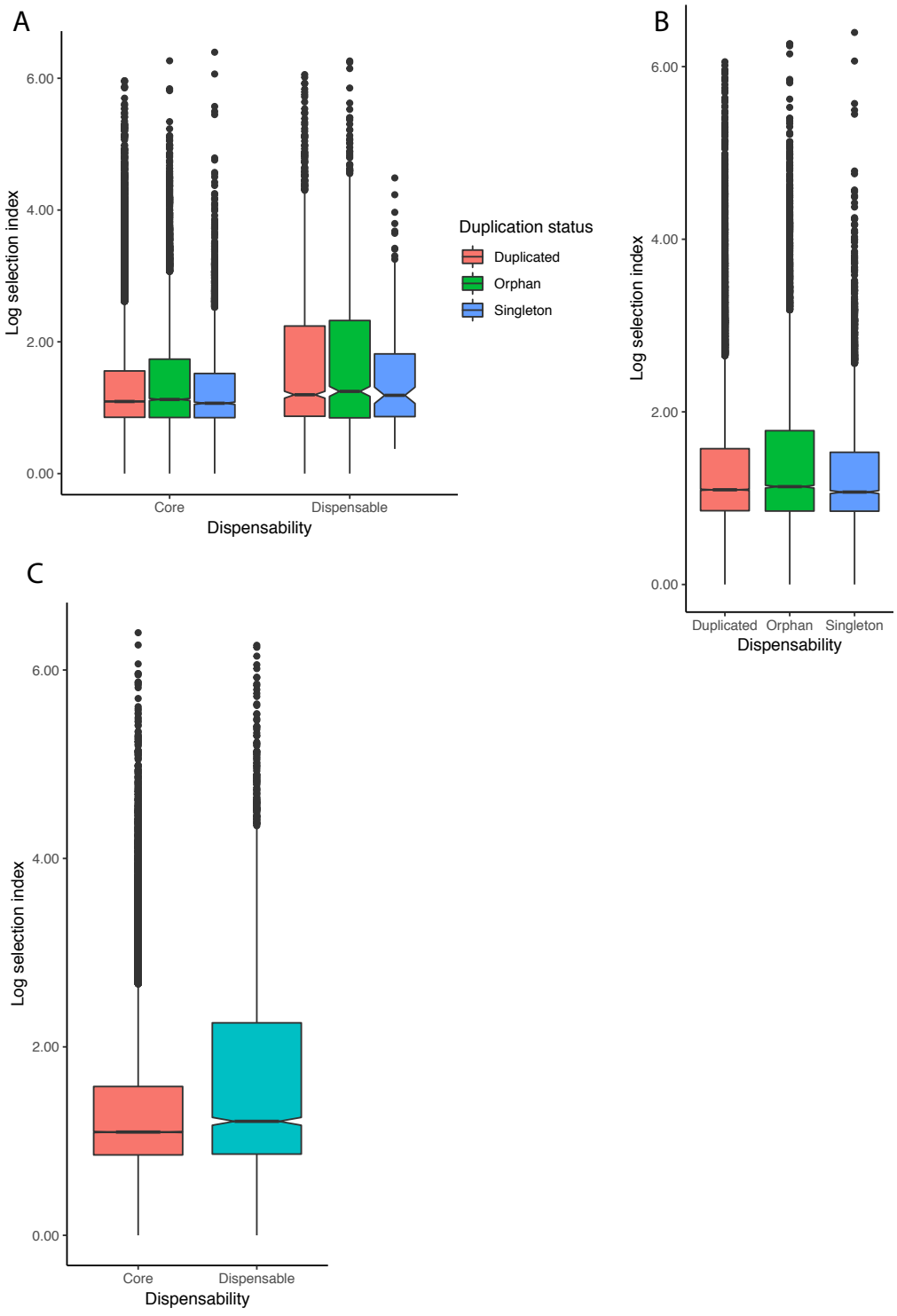

Figure 5.8. Selection indices of genes in each duplicated or dispensable category. A) Both dispensability and duplication status together, B) duplication status only, and C) dispensability only.

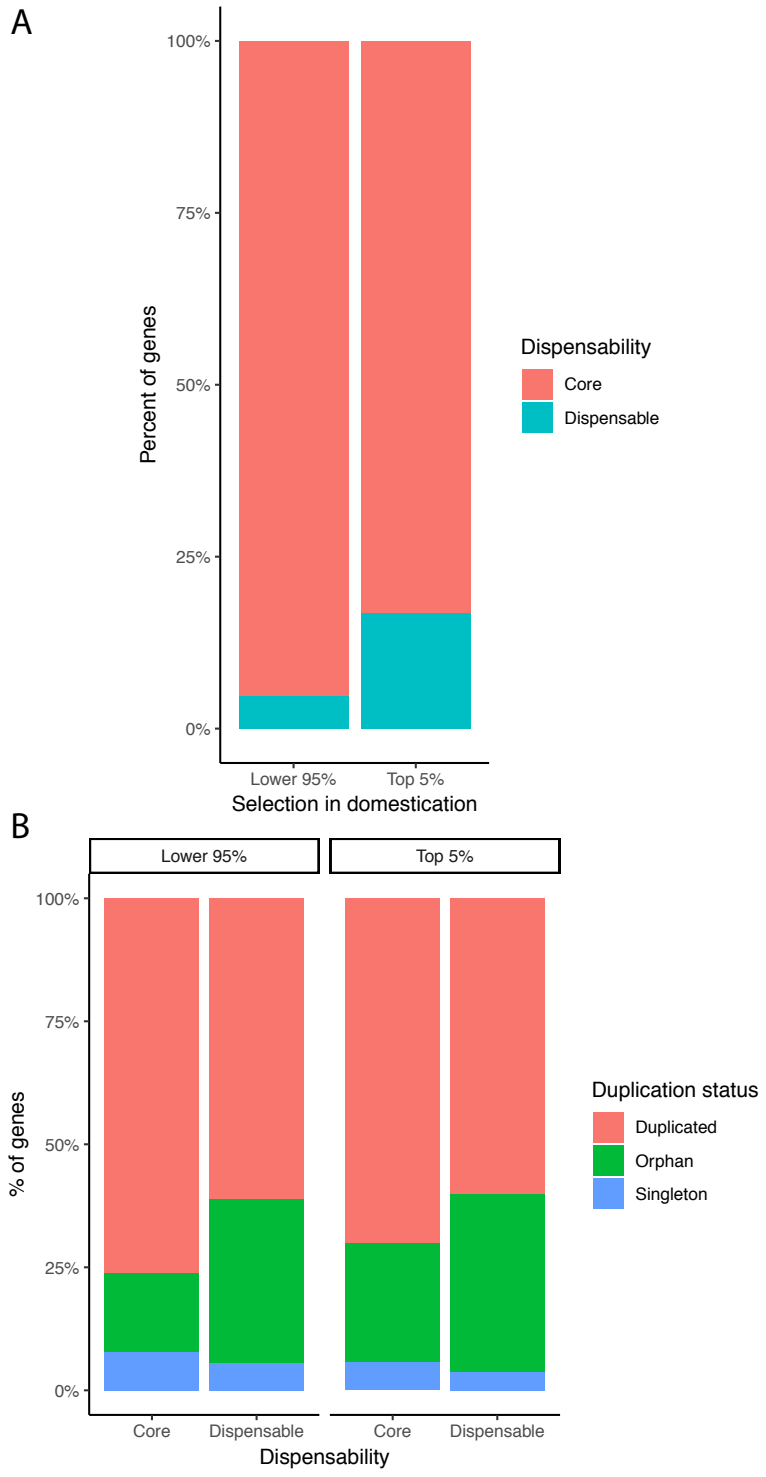

Figure 5.9. Comparing dispensability and duplication status among genes in the top 5% of π_w/π_d. A) Comparing dispensability only, B) comparing dispensability and duplication status together.

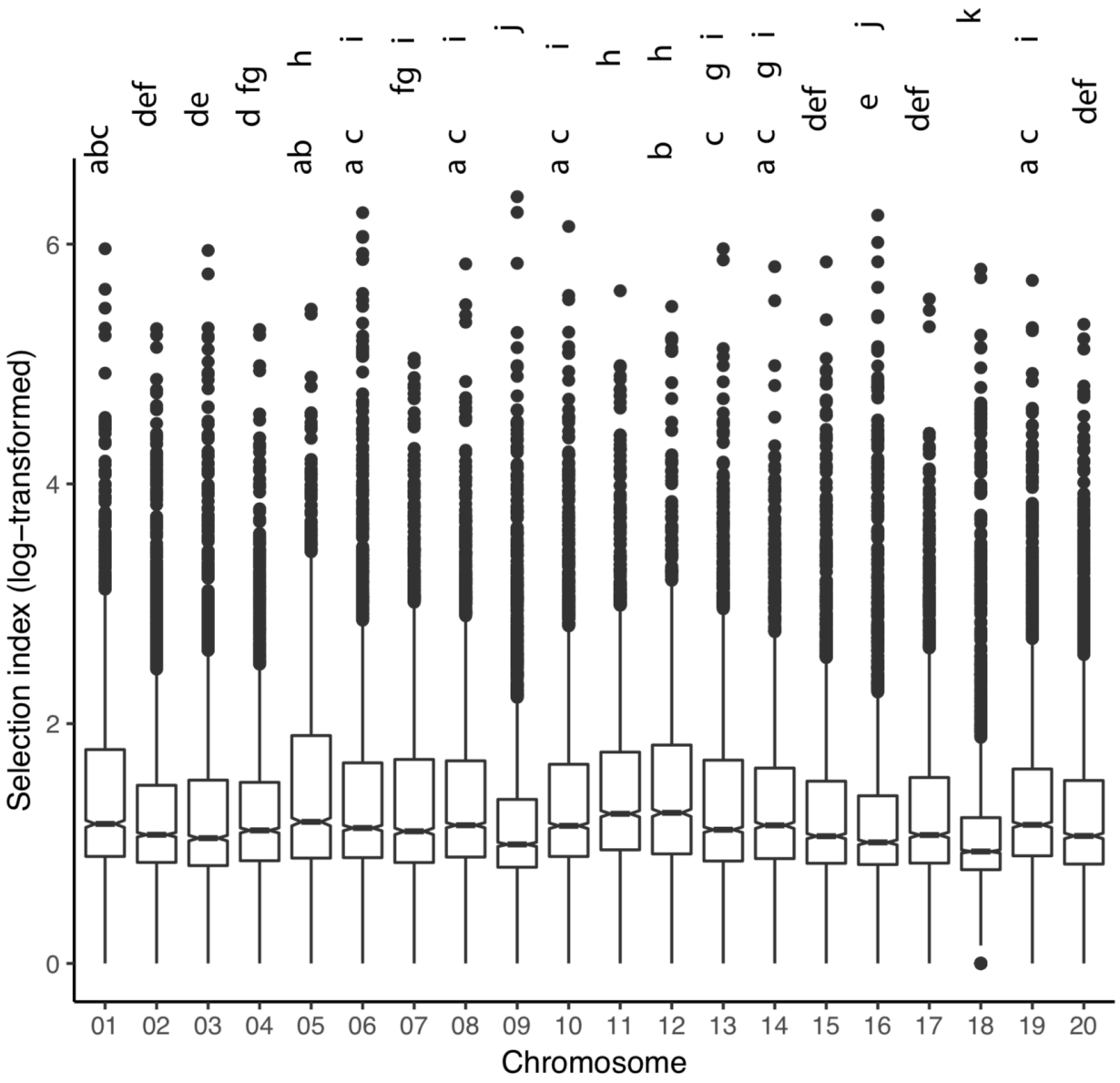

Fig 5.S1. Distribution of selection indices across the 20 chromosomes of soybean.

Grouping letters at top are based on pairwise Wilcoxon Rank Sum tests.

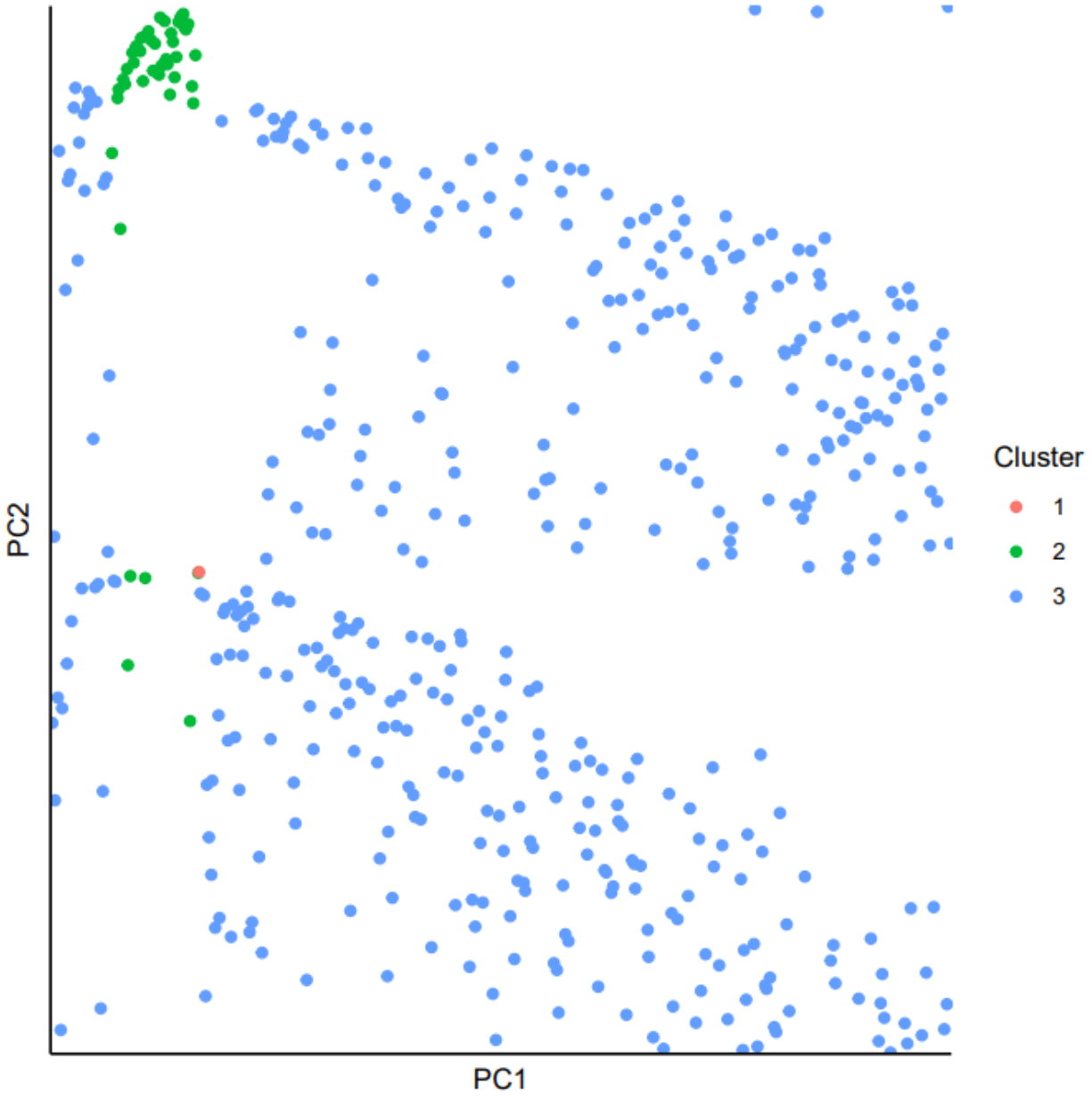

Figure 5.S2. K-means cluster analysis of PCA data from Fig 5.1 with k=3 (nstart=25, iterations = 1000). Cluster 1 has only one member, and cluster 2 includes all of the *G. soja* lines but also a few domesticated *G. max* lines as well.

CHAPTER 6

CONCLUSIONS

Polyploidy and gene duplication have been increasingly appreciated as central players in the evolution of the genomes of nearly all life on earth. While gene flow, genetic drift, and selection are the primary forces that drive evolution, mutation is still the ultimate source of all genetic variation. Duplication of DNA, whether through meiotic errors, replication slippage, unequal crossover, transposon activity, or otherwise has been undervalued until recently as an indispensable source of variation in natural populations. Duplication of genes can allow for one resultant copy to explore more of the available evolutionary space and take on new functions, driving phenotypic diversity and adaptation. It can also create entirely new species, such as when two divergent genomes come together in one nucleus to create a new allopolyploid species that is effectively reproductively isolated from either of its progenitors.

The work presented in this dissertation present new and re-examined evidence of the importance of gene and genome duplication in legumes, and especially in soybean. Soybean has a large genome with over 55,000 genes and 20 chromosomes despite being diploid. This is probably due to the multiple paleopolyploidy events in its history. The first part of this work which examined WGD and segmental duplications demonstrated that the more recent blocks of synteny arising from the *Glycine*-specific duplication contained more genes in longer blocks. They also contained paralogous gene pairs with highly similar expression patterns across tissues, suggesting that these more recent WGD duplicate pairs of genes have maintained functional similarity despite up to 13 million years passing since the most recent WGD in soybean. Furthermore, these more recent gene pairs also had maintained the same methylation status more

often than the older duplicated pairs, as shown in the abundance of gene pairs which both were CG gene body methylated among newer *Glycine* pairs but not among older duplicated pairs, which had more unmethylated + CG pairs. These results indicate there is a tendency for soybean's most recent duplicated pairs to be retained together, expressed together, and maintain the same methylation status together more than their older counterparts (compared to the other 2 duplications detected in soybean), even when accounting for the age of the duplications.

Further examination of the most recent duplication event in soybean was warranted by the conclusion that the more recent soybean duplicate gene pairs are similar in their expression, methylation, and retention levels. Thus, an algorithm (TetrAssign) was developed to reconstruct and compare ancient tetraploid subgenomes in diploidized species like maize and soybean. The results indicated that soybean's tetraploid subgenomes were more highly rearranged and had less bias in gene deletion, expression, or methylation, than maize's. Under the expectations set by the genome dominance hypothesis, this might mean that soybean's ancient tetraploid subgenomes experienced little to no genome dominance post-polyploidy and that perhaps this most recent tetraploidy in soybean was more like a segmental allopolyploidy or autopolyploidy, or that the extinct progenitor species which hybridized to form *Glycine* were much more closely related, than was previously thought.

Soybean appears to be something of an outlier among its crop legume cousins with its large, duplicated, diploid genome with over 55,000 genes (with most other diploid legumes having about 35,000 or fewer). While the existence of a *Glycine*-specific paleotetraploidy was proposed early on, it was unclear if this was the primary driver of the size and redundancy of the soybean genome. To examine this question, a comprehensive set of orthogroups was built for soybean and several other legume genomes. This revealed that while legume genes were

commonly in large gene families, duplicate gene copies were less likely to be deleted in the *Glycine* clade (though, interestingly, this was also true of the *Arachis* clade, despite those diploid species having 41,840 genes at maximum). Furthermore, the maximum likelihood approach used here found that duplicates arising from the *Glycine* were dramatically more likely to be retained post-diploidization than those from the legume-shared *Faboideae* event. Both results indicate that while *Glycine* may not be alone in retaining many duplicate gene copies after millions of years, its especially high level of duplicate retention and large gene family size is most likely being driven by the nature of the *Glycine* tetraploidy event. Again, this may be because this event involved two very similar genomes forming a tetraploid, leading to little genome dominance and a strong dosage balance effect acting across the soybean genome.

Duplicated genes and genomes can evolve in many ways over millions of years, giving rise to substantial variation in genetic content between species; but these processes did not halt when humans began to cultivate crop plants. Indeed, many results indicate that the evolutionary processes acting upon duplicate genes are still ongoing today, perhaps driving structural variation in the germplasm of the plants most important to humans. However, this understanding has yet to be fully harnessed by plant breeders. The reasons for this are likely numerous, but one possible contributor is that information on how exactly duplicated genes or genomes could affect breeding efforts is quite sparse. Currently, little is known how variable duplicate genes are within germplasm (i.e. how many gene families vary in their size or membership within a species), or how strongly structural variation (e.g. presence-absence variation or copy number variation) is linked to common SNPs or other markers. Still, it is possible that a focus on point variations like SNPs has meant many important structural variations, like variation in gene family size,

presence-absence variation, copy number variation, or the presence of novel (i.e. orphan) genes have gone unnoticed.

While the first generation of genome assemblies and annotations for major crop plants like maize, soybean, cotton, tomato, and more have enabled more precise breeding or trait mapping and a better understanding of the architecture and functional genetics of important traits, much work remains to be done in understanding the true breadth of genomic diversity available in our crops' germplasm. Reference sequences are limited by nature, however. If other genotypes in the germplasm of a plant do not share the same gene set as the reference genome, potentially valuable variation is missed due to a reference-based approach to defining diversity. This is because aligning sequences from diverse genotypes to a reference sequence means novel sequences in the diverse genotypes are thrown out, as they cannot be aligned. Describing the kind of diversity represented by these discarded sequences requires an entirely different approach to understanding population genetics, one largely incompatible with the outdated "beads-on-a-string" model of genomic comparison, where every genome contains the same loci but with varying alleles.

With the release of the first few *de novo* assemblies of diverse genotypes for important crop plants like PH207 maize, a clearer picture of the true genomic variation present in plants is coming into view: that of the pan-genome. Much of the genic variation that has been revealed in these first few studies of pan-genomes have shown that many of the genes not shared between genotypes are members of dynamic gene families, and that these families' membership and size could be driving potentially important traits for breeders like disease resistance or stress responses. With newer long-range sequencing technologies driving down the cost of *de novo*

assembly of diverse genotypes, a new understanding of pan-genomic variation and how it has been driven by the evolution of duplicate genes and genome segments is certainly on the horizon.

This work has demonstrated that soybean's genome is highly duplicated despite millions of years passing since its most recent paleopolyploidy through syntenic alignments of genome segments and comparing those segments' expression and methylation patterns. It has also shown that soybean's ancient tetraploid subgenomes were perhaps quite similar in their expression, methylation, and gene retention via reconstruction of these subgenomes and comparing them to maize. A phylogenetic analysis of soybean's genes along with those of other legumes showed that the gene deletion and duplication rates along the legume species tree were variable. It also suggested that duplicate gene evolution among these legumes was mostly driven by polyploidy, and that soybean's most recent tetraploidy had significantly higher duplicate gene retention rates than the older legume-shared duplication, raising questions about the supposed allotetraploid origin of soybean. Finally, this work showed that there is considerable presence-absence variation among soybean's elite and wild germplasm, and that dispensable genes (genes missing in one or more soybean accessions) and/or orphan genes (genes with no orthologs or paralogs, and probably the results of ancient duplications) may have been key players in the domestication of soybean. In all, these results indicate that understanding soybean's history of duplication has affected not only its evolution over millions of years but has also possibly affected its use as a crop plant in the past few thousand years. It still remains to be determined whether structural variations or novel genes in varying accessions are tightly associated with common SNPs or other markers. However, if there is weak linkage between traditional markers and these novel duplication-derived sequences, breeders may be able to use these variants to identify novel genes

or gene families that affect valuable traits like disease resistance and harness them to improve the crop and make greater genetic gains.

While polyploidy, duplication, and duplicate gene evolution pose interesting questions from a basic research standpoint, much work remains to be done in connecting this knowledge base to making measurable gains in breeding. With new technologies, new techniques, and new knowledge of genome evolution, perhaps breeders of the future will be able to absorb and harness this knowledge into new breeding pipelines, new trait discoveries, and ultimately new and improved crop varieties suited to the task of feeding more humans in a rapidly changing world.

REFERENCES

Abel, S., and Becker, H. (2007). The effect of autopolyploidy on biomass production in homozygous lines of Brassica rapa and Brassica oleracea. *Plant breeding* **126**, 642-643.

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current opinion in plant biology* **8**, 135-141.

Adolph, K. W. (1991). "Advanced Techniques in Chromosome Research," Taylor & Francis.

Akhunov, E. D., Akhunova, A. R., and Dvorak, J. (2007). Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. *Mol Biol Evol* **24**, 539-50.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363.

Anderson, L. K., Lai, A., Stack, S. M., Rizzon, C., and Gaut, B. S. (2006). Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome research* **16**, 115-122.

Andrews, M., and Andrews, M. E. (2017). Specificity in Legume-Rhizobia Symbioses. *International Journal of Molecular Sciences* **18**, 705.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J. F., Sabot, F., Kudrna, D., Ammiraju, J. S. S., Schuster, S. C., Carlson, J. E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Bérard, A., Viot, C., Boccara, M., Risterucci, A. M., Guignon, V., Sabau, X., Axtell, M. J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J. M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W. R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., and Lanaud, C. (2010). The genome of Theobroma cacao. *Nature Genetics* **43**, 101.

Armstrong, J. M. (1954). CYTOLOGICAL STUDIES IN ALFALFA POLYPLOIDS. *Canadian Journal of Botany* **32**, 531-542.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25.

Assis, R., and Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila. *Proceedings of the National Academy of Sciences* **110**, 17409-17414.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology* **14**, 283-291.

Baptista, A., Pinho, O., Pinto, E., Casal, S., Mota, C., and Ferreira, I. M. P. L. V. O. (2017). Characterization of protein and fat composition of seeds from common beans (Phaseolus vulgaris L.), cowpea (Vigna unguiculata L. Walp) and bambara groundnuts (Vigna subterranea L. Verdc) from Mozambique. *Journal of Food Measurement and Characterization* **11**, 442-450.

Barcaccia, G., Meneghetti, S., Albertini, E., Triest, L., and Lucchin, M. (2003). Linkage mapping in tetraploid willows: segregation of molecular markers and estimation of linkage phases support an allotetraploid structure for Salixalba× Salix fragilis interspecific hybrids. *Heredity* **90**, 169.

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* **210**, 391-398.

Belling, J. (1912). "Collected Reprints."

Bellora, N., Toll-Riera, M., Mar Albà, M., Castelo, R., Estivill, X., Armengol, L., and Bosch, N. (2008). Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Molecular Biology and Evolution* **26**, 603-612.

Bernardo, R. (2016). Bandwagons I, too, have known.

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M. C., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C. E., Umale, P., Araújo, A. C. G., Kozik, A., Do Kim, K., Burow, M. D., Varshney, R. K., Wang, X., Zhang, X., Barkley, N., Guimarães, P. M., Isobe, S., Guo, B., Liao, B., Stalker, H. T., Schmitz, R. J., Scheffler, B. E., Leal-Bertioli, S. C. M., Xun, X., Jackson, S. A., Michelmore, R., and Ozias-Akins, P. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genetics* **48**, 438.

Bewick, A. J., Ji, L., Niederhuth, C. E., Willing, E.-M., Hofmeister, B. T., Shi, X., Wang, L., Lu, Z., Rohr, N. A., and Hartwig, B. (2016). On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Sciences* **113**, 9111-9116.

Bewick, A. J., and Schmitz, R. J. (2017). Gene body DNA methylation in plants. *Current opinion in plant biology* **36**, 103-110.

Bhakta, M. S., Jones, V. A., and Vallejos, C. E. (2015). Punctuated Distribution of Recombination Hotspots and Demarcation of Pericentromeric Regions in Phaseolus vulgaris L. *PLOS ONE* **10**, e0116822.

Bingham, E. T., and Gillies, C. B. (1971). CHROMOSOME PAIRING, FERTILITY, AND CROSSING BEHAVIOR OF HAPLOIDS OF TETRAPLOID ALFALFA, MEDICAGO SATIV A L. *Canadian Journal of Genetics and Cytology* **13**, 195-202.

Birchler, J. A., and Newton, K. J. (1981). MODULATION OF PROTEIN LEVELS IN CHROMOSOMAL DOSAGE SERIES OF MAIZE: THE BIOCHEMICAL BASIS OF ANEUPLOID SYNDROMES. *Genetics* **99**, 247-266.

Birchler, J. A., and Veitia, R. A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *The Plant Cell* **19**, 395-402.

Birchler, J. A., and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 14746-14753.

Birchler, J. A., and Veitia, R. A. (2014). The Gene Balance Hypothesis: dosage effects in plants. *Methods Mol Biol* **1112**, 25-32.

Blakeslee, A. F., and Belling, J. (1924). Chromosomal mutations in the Jimson weed, Datura stramonium. *Journal of Heredity* **15**, 195-206.
Blakeslee, A. F., Belling, J., and Farnham, M. E. (1923). Inheritance in Tetraploid Daturas. *Botanical Gazette* **76**, 329-373.
Blanc, G., and Wolfe, K. H. (2004). Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *The Plant Cell* **16**, 1679-1691.
Bolger, M., Schwacke, R., Gundlach, H., Schmutzer, T., Chen, J., Arend, D., Oppermann, M., Weise, S., Lange, M., Fiorani, F., Spannagl, M., Scholz, U., Mayer, K., and Usadel, B. (2017). From plant genomes to phenotypes. *Journal of Biotechnology* **261**, 46-52.
Bolot, S., Abrouk, M., Masood-Quraishi, U., Stein, N., Messing, J., Feuillet, C., and Salse, J. (2009). The 'inner circle' of the cereal genomes. *Current Opinion in Plant Biology* **12**, 119-125.
Bonierbale, M. W., Plaisted, R. L., and Tanksley, S. D. (1988). RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics* **120**, 1095-1103.
Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2015). The Double-Reduction Landscape in Tetraploid Potato as Revealed by a High-Density Linkage Map. *Genetics* **201**, 853-863.
Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433-8.
Bowman, B. H., and Kurosky, A. (1982). Haptoglobin: the evolutionary product of duplication, unequal crossing over, and point mutation. *In* "Advances in human genetics", pp. 189-261. Springer.
Brea, M., Zamuner, A. B., Matheos, S. D., Iglesias, A., and Zucol, A. F. (2008). Fossil wood of the Mimosoideae from the early Paleocene of Patagonia, Argentina. *Alcheringa: An Australasian Journal of Palaeontology* **32**, 427-441.
Brown, C. R. (1993). Outcrossing rate in cultivated autotetraploid potato. *American Potato Journal* **70**, 725-734.
Brownfield, L., and Köhler, C. (2011). Unreduced gamete formation in plants: mechanisms and prospects. *Journal of Experimental Botany* **62**, 1659-1668.
Brummer, E. C., Cazcarro, P. M., and Luth, D. (1999). Ploidy determination of alfalfa germplasm accessions using flow cytometry. *Crop science* **39**, 1202-1207.
Bruneau, A., Mercure, M., Lewis, G. P., and Herendeen, P. S. (2008). Phylogenetic patterns and diversification in the caesalpinioid legumesThis paper is one of a selection of papers published in the Special Issue on Systematics Research. *Botany* **86**, 697-718.
Buckler, E. S., Thornsberry, J. M., and Kresovich, S. (2001). Molecular diversity, structure and domestication of grasses. *Genetics Research* **77**, 213-218.
Buggs, R. J. A., Wendel, J. F., Doyle, J. J., Soltis, D. E., Soltis, P. S., and Coate, J. E. (2014). The legacy of diploid progenitors in allopolyploid gene expression patterns. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**.
Busbice, T. H., and Wilsie, C. P. (1966). Inbreeding depression and heterosis in autotetraploids with application to Medicago sativa L. *Euphytica* **15**, 52-67.
Butler, M. A., and King, A. A. (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist* **164**, 683-695.

Byrd, A. K., and Raney, K. D. (2017). Structure and function of Pif1 helicase. *Biochemical Society transactions* **45**, 1159-1171.
Byrne, K. P., and Wolfe, K. H. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175**, 1341-1350.
Byrnes, J. K., Morris, G. P., and Li, W.-H. (2006). Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Molecular biology and evolution* **23**, 1136-1143.
Cai, H., and Morishima, H. (2002). QTL clusters reflect character associations in wild and cultivated rice. *Theoretical and Applied Genetics* **104**, 1217-1228.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421-421.
Cannon, S. B., Ilut, D., Farmer, A. D., Maki, S. L., May, G. D., Singer, S. R., and Doyle, J. J. (2010). Polyploidy Did Not Predate the Evolution of Nodulation in All Legumes. *PLOS ONE* **5**, e11630.
Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biology* **4**, 10.
Cannon, S. B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., and Schiex, T. (2006). Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proceedings of the National Academy of Sciences* **103**, 14959-14964.
Casola, C., and Lawing, A. M. (2018). The nonrandom evolution of gene families. *American Journal of Botany*.
Ceccarelli, S., Grando, S., Maatougui, M., Michael, M., Slash, M., Haghparast, R., Rahmanian, M., Taheri, A., Al-Yassin, A., and Benbelkacem, A. (2010). Plant breeding and climate changes. *The Journal of Agricultural Science* **148**, 627-637.
Chang, P. L., Dilkes, B. P., McMahon, M., Comai, L., and Nuzhdin, S. V. (2010). Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biol* **11**, R125.
Chen, P. D., Tsujimoto, H., and Gill, B. S. (1994). Transfer of PhI genes promoting homoeologous pairing from Triticum speltoides to common wheat. *Theoretical and Applied Genetics* **88**, 97-101.
Chen, Z. J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* **58**, 377-406.
Cheng, F., Mandáková, T., Wu, J., Xie, Q., Lysak, M. A., and Wang, X. (2013). Deciphering the diploid ancestral genome of the mesohexaploid Brassica rapa. *The Plant Cell*, tpc. 113.110486.
Cheng, F., Sun, C., Wu, J., Schnable, J., Woodhouse, M. R., Liang, J., Cai, C., Freeling, M., and Wang, X. (2016). Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. *New Phytol* **211**, 288-99.
Cheptou, P. O. (2012). Clarifying Baker's Law. *Annals of Botany* **109**, 633-641.
Cliften, P. F., Fulton, R. S., Wilson, R. K., and Johnston, M. (2006). After the duplication: gene loss and adaptation in Saccharomyces genomes. *Genetics* **172**, 863-872.
Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836-46.

Contreras-Soto, R. I., De Oliveira, M. B., Costenaro-Da-Silva, D., Scapim, C. A., and Schuster, I. (2017). Population structure, genetic relatedness and linkage disequilibrium blocks in cultivars of tropical soybean (Glycine max). *Euphytica* **213**.

Crow, K. D., and Wagner, G. P. (2005). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular biology and evolution* **23**, 887-892.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H., and dePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**, 738-49.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., Alberti, A., Bernard, M., Correa, M., Ayyampalayam, S., McKain, M. R., Leebens-Mack, J., Burgess, D., Freeling, M., Mbéguié-A-Mbéguié, D., Chabannes, M., Wicker, T., Panaud, O., Barbosa, J., Hribova, E., Heslop-Harrison, P., Habas, R., Rivallan, R., Francois, P., Poiron, C., Kilian, A., Burthia, D., Jenny, C., Bakry, F., Brown, S., Guignon, V., Kema, G., Dita, M., Waalwijk, C., Joseph, S., Dievart, A., Jaillon, O., Leclercq, J., Argout, X., Lyons, E., Almeida, A., Jeridi, M., Dolezel, J., Roux, N., Risterucci, A.-M., Weissenbach, J., Ruiz, M., Glaszmann, J.-C., Quétier, F., Yahiaoui, N., and Wincker, P. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213.

Dagan, T., Martin, W., Kilian, B., Özkan, H., Salamini, F., Walther, A., and Kohl, J. (2007). Molecular Diversity at 18 Loci in 321 Wild and 92 Domesticate Lines Reveal No Reduction of Nucleotide Diversity during Triticum monococcum (Einkorn) Domestication: Implications for the Origin of Agriculture. *Molecular Biology and Evolution* **24**, 2657-2668.

Dandia, S., Dhar, A., and Sengupta, K. (1990). Meiosis in natural decosaploid (22X) Morus nigra L. *Cytologia* **55**, 505-509.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**, 2156-2158.

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271.

De Bodt, S., Maere, S., and Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends Ecol Evol* **20**, 591-7.

De Faria, S., Lewis, G., Sprent, J., and Sutherland, J. (1989). Occurrence of nodulation in the Leguminosae. *New Phytologist* **111**, 607-619.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences* **110**, 2898-2903.

De Storme, N., and Mason, A. (2014). Plant speciation through chromosome instability and ploidy change: Cellular mechanisms, molecular factors and evolutionary relevance. *Current Plant Biology* **1**, 10-33.

Delgado, C. L. (2003). Rising consumption of meat and milk in developing countries has created a new food revolution. *The Journal of nutrition* **133**, 3907S-3910S.

DeSA, U. (2013). World population prospects: the 2012 revision. *Population division of the department of economic and social affairs of the United Nations Secretariat, New York*.
Desai, S. S. (1997). Down syndrome: a review of the literature. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontics* **84**, 279-285.
Devos, K. M. (2005). Updating the 'crop circle'. *Current opinion in plant biology* **8**, 155-162.
DeVries, H. (1915). The Coefficient of Mutation in Oenothera biennis L. *Botanical Gazette* **59**, 169-196.
Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21.
Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The Molecular Genetics of Crop Domestication. *Cell* **127**, 1309-1321.
Domazet-Loso, T., and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Res* **13**, 2213-9.
Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., and Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC evolutionary biology* **11**, 47.
Doyle, J. J., Doyle, J. L., Rauscher, J. T., and Brown, A. H. D. (2004). Evolution of the perennial soybean polyploid complex (Glycine subgenus Glycine): a study of contrasts. *Biological Journal of the Linnean Society* **82**, 583-597.
Doyle, J. J., and Luckow, M. A. (2003). The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiology* **131**, 900-910.
Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**, 1862-1866.
Dvorak, J., and Lukaszewski, A. J. (2000). Centromere association is an unlikely mechanism by which the wheat Ph1 locus regulates metaphase I chromosome pairing between homoeologous chromosomes. *Chromosoma* **109**, 410-414.
Edger, P. P., McKain, M. R., Bird, K. A., and VanBuren, R. (2018). Subgenome assignment in allopolyploids: challenges and future directions. *Current Opinion in Plant Biology* **42**, 76-80.
Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., Bewick, A. J., Ji, L., Platts, A. E., Bowman, M. J., Childs, K. L., Washburn, J. D., Schmitz, R. J., Smith, G. D., Pires, J. C., and Puzey, J. R. (2017). Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *Plant Cell* **29**, 2150-2167.
El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C., and Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (Triticum aestivum). *New Phytol* **213**, 1477-1486.
Emery, M., Willis, M. M. S., Hao, Y., Barry, K., Oakgrove, K., Peng, Y., Schmutz, J., Lyons, E., Pires, J. C., Edger, P. P., and Conant, G. C. (2018). Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genet* **14**, e1007267.
Emms, D. M., Covshoff, S., Hibberd, J. M., and Kelly, S. (2016). Independent and Parallel Evolution of New Genes by Gene Duplication in Two Origins of C4 Photosynthesis Provides New Insight into the Mechanism of Phloem Loading in C4 Species. *Molecular Biology and Evolution* **33**, 1796-1806.

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157.
Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L., and Gaut, B. S. (1998). Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences* **95**, 4441-4446.
FAO (2013). Feeding nine billion in 2050. *In* "UN FAO".
Faria, S. d., Lewis, G., Sprent, J., and Sutherland, J. (1989). Occurrence of nodulation in the Leguminosae. *New Phytologist* **111**, 607-619.
Fawcett, J. A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences* **106**, 5737-5742.
Flagel, L., Udall, J., Nettleton, D., and Wendel, J. (2008). Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC biology* **6**, 16.
Flagel, L. E., and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**, 557-564.
Flagel, L. E., and Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist* **186**, 184-193.
Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics* **151**, 1531-1545.
Freeling, M., Scanlon, M. J., and Fowler, J. E. (2015). Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Current Opinion in Genetics & Development* **35**, 110-118.
Freeling, M., Woodhouse, M. R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J. C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* **15**, 131-139.
Friedman, N., Cai, L., and Xie, X. S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* **97**, 168302.
Friedman, W. E. (2009). The meaning of Darwin's “abominable mystery”. *American Journal of Botany* **96**, 5-21.
Furman, B. L. S., Dang, U. J., Evans, B. J., and Golding, G. B. (2018). Divergent subgenome evolution after allopolyploidization in African clawed frogs (Xenopus). *Journal of Evolutionary Biology* **31**, 1945-1958.
Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A., and Köhler, N. (1999). Discovery of tetraploidy in a mammal. *Nature* **401**, 341-341.
Galloway, L., Etterson, J., and Hamrick, J. (2003). Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, Campanula americana. *Heredity* **90**, 308.
Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D’Hont, A., and Freeling, M. (2014). Two Evolutionarily Distinct Classes of Paleopolyploidy. *Molecular Biology and Evolution* **31**, 448-454.
Gaut, B. S. (2001). Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res* **11**, 55-66.

Gaut, B. S., and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences* **94**, 6809-6814.
Gay, D. M. (1990). Usage summary for selected optimization routines.
Gehring, M., and Henikoff, S. (2007). DNA methylation dynamics in plant genomes. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1769**, 276-286.
Ghenu, A.-H., Bolker, B. M., Melnick, D. J., and Evans, B. J. (2016). Multicopy gene family evolution on primate Y chromosomes. *BMC genomics* **17**, 157-157.
Gibson, T. J., and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends in genetics* **14**, 46-49.
Gill, N., Findley, S., Walling, J. G., Hans, C., Ma, J., Doyle, J., Stacey, G., and Jackson, S. A. (2009). Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant physiology* **151**, 1167-1174.
Gizlice, Z., Carter, T. E., and Burton, J. W. (1994). Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. **34**, 1143-1151.
Gonzalo, A., Lucas, M.-O., Charpentier, C., Lloyd, A. H., and Jenczewski, E. (2018). Meiotic effects of MSH4 copy number variation support an adaptive role for post-polyploidy gene loss. *bioRxiv*, 482521.
Goodman, M. M., Stuber, C. W., Newton, K., and Weissinger, H. H. (1980). Linkage relationships of 19 enzyme Loci in maize. *Genetics* **96**, 697-710.
Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**, D1178-D1186.
Gottlieb, L. (1982). Conservation and duplication of isozymes in plants. *Science* **216**, 373-380.
Graham, P. H., and Vance, C. P. (2003). Legumes: Importance and Constraints to Greater Use. *Plant Physiology* **131**, 872-877.
Griffiths, S., Sharp, R., Foote, T. N., Bertin, I., Wanous, M., Reader, S., Colas, I., and Moore, G. (2006). Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439**, 749.
Grigg, D. (1995). The pattern of world protein consumption. *Geoforum* **26**, 1-17.
Gross, B. L., and Olsen, K. M. (2010). Genetic perspectives on crop domestication. *Trends in plant science* **15**, 529-537.
Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., and Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63.
Guo, H., Wang, X., Gundlach, H., Mayer, K. F. X., Peterson, D. G., Scheffler, B. E., Chee, P. W., and Paterson, A. H. (2014). Extensive and Biased Intergenomic Nonreciprocal DNA Exchanges Shaped a Nascent Polyploid Genome, Gossypium (Cotton). *Genetics* **197**, 1153-1163.
Haas, B. J., Delcher, A. L., Wortman, J. R., and Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-6.
Hahn, M. W., Demuth, J. P., and Han, S.-G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics*.
Hammer, K. (1984). Das Domestikationssyndrom. *Die Kulturpflanze* **32**, 11-34.
Hamrick, J. L., and Godt, M. W. (1996). Effects of life history traits on genetic diversity in plant species. *Phil. Trans. R. Soc. Lond. B* **351**, 1291-1298.

Han, Y., Zheng, D., Vimolmangkang, S., Khan, M. A., Beever, J. E., and Korban, S. S. (2011). Integration of physical and genetic maps in apple confirms whole-genome and segmental duplications in the apple genome. *Journal of Experimental Botany* **62**, 5117-5130.

Hanelt, P. (1986). Pathways of Domestication with Regard to Crop Types (Grain Legumes, Vegetables). Vol. 16, pp. 179-199.

Hargrove, M. S., Barry, J. K., Brucker, E. A., Berry, M. B., Phillips, G. N., Jr., Olson, J. S., Arredondo-Peter, R., Dean, J. M., Klucas, R. V., and Sarath, G. (1997). Characterization of recombinant soybean leghemoglobin a and apolar distal histidine mutants. *J Mol Biol* **266**, 1032-42.

Harlan, J. R., and deWet, J. M. J. (1975). On Ö. Winge and a Prayer: The origins of polyploidy. *The Botanical Review* **41**, 361-390.

Hartl, D. L., Clark, A. G., and Clark, A. G. (1997). "Principles of population genetics," Sinauer associates Sunderland, MA.

He, Z., Zhai, W., Wen, H., Tang, T., Wang, Y., Lu, X., Greenberg, A. J., Hudson, R. R., Wu, C.-I., and Shi, S. (2011). Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS genetics* **7**, e1002100.

Herrera, J., Combes, M.-C., Anthony, F., Charrier, A., and Lashermes, P. (2002). Introgression into the allotetraploid coffee (Coffea arabica L.): segregation and recombination of the C. canephora genome in the tetraploid interspecific hybrid (C. arabica× C. canephora). *Theoretical and Applied Genetics* **104**, 661-668.

Heyn, F. W. (1977). Analysis of unreduced gametes in the Brassiceae by crosses between species and ploidy levels. *Zeitschrift für Pflanzenzüchtung*.

Hiesey, W. M. (1966). The Genetics of Colonizing Species. Proceedings of the First International Union of Biological Sciences Symposia on General Biology. H. G. Baker , G. Ledyard Stebbins. *The Quarterly Review of Biology* **41**, 418-419.

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X., and Frazer, K. A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics* **38**, 82-85.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell* **26**, 121-135.

Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A. G., Fields, C. J., Wright, C. L., Koehler, K., Springer, N. M., Buckler, E., Buell, C. R., de Leon, N., Kaeppler, S. M., Childs, K. L., and Mikel, M. A. (2016). Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. **28**, 2700-2714.

Hollister, J. D., Smith, L. M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B. S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2322-2327.

Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. (1998). Dissecting the Regulatory Circuitry of a Eukaryotic Genome. *Cell* **95**, 717-728.

Hougaard, B. K., Madsen, L. H., Sandal, N., De Carvalho Moretzsohn, M., Fredslund, J., Schauser, L., Nielsen, A. M., Rohde, T., Sato, S., Tabata, S., Bertioli, D. J., and Stougaard, J. (2008). Legume Anchor Markers Link Syntenic Regions Between

Phaseolus vulgaris, Lotus japonicus, Medicago truncatula and Arachis. *Genetics* **179**, 2299-2312.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., and Cregan, P. B. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences* **103**, 16666-16671.

Innes, R. W., Ameline-Torregrosa, C., Ashfield, T., Cannon, E., Cannon, S. B., Chacko, B., Chen, N. W. G., Couloux, A., Dalwani, A., Denny, R., Deshpande, S., Egan, A. N., Glover, N., Hans, C. S., Howell, S., Ilut, D., Jackson, S., Lai, H., Mammadov, J., del Campo, S. M., Metcalf, M., Nguyen, A., O'Bleness, M., Pfeil, B. E., Podicheti, R., Ratnaparkhe, M. B., Samain, S., Sanders, I., Ségurens, B., Sévignac, M., Sherman-Broyles, S., Thareau, V., Tucker, D. M., Walling, J., Wawrzynski, A., Yi, J., Doyle, J. J., Geffroy, V., Roe, B. A., Maroof, M. A. S., and Young, N. D. (2008). Differential Accumulation of Retroelements and Diversification of NB-LRR Disease Resistance Genes in Duplicated Regions following Polyploidy in the Ancestor of Soybean. *Plant Physiology* **148**, 1740-1759.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* **436**, 793.

Ising, G. (1966). Cytogenetic studies in Cyrtanthus I. Segregation in an allotetraploid. *Hereditas* **56**, 27-53.

Jackson, R. C. (1976). Evolution and systematic significance of polyploidy. *Annual Review of Ecology and Systematics* **7**, 209-234.

Jackson, R. C. (1982). Polyploidy and Diploidy: New Perspectives on Chromosome Pairing and Its Evolutionary Implications. *American Journal of Botany* **69**, 1512-1523.

Jelesko, J. G., Harper, R., Furuya, M., and Gruissem, W. (1999). Rare germinal unequal crossing-over leading to recombinant gene formation and gene duplication in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **96**, 10302-10307.

Jiao, Y., and Paterson, A. H. (2014). Polyploidy-associated genome modifications during land plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., and Soltis, P. S. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-40.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S.-i., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., and Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714.

Kawashima, K., Suganuma, N., Tamaoki, M., and Kouchi, H. (2001). Two types of pea leghemoglobin genes showing different O2-binding affinities and distinct patterns of spatial expression in nodules. *Plant physiology* **125**, 641-651.

Keller, B., and Feuillet, C. (2000). Colinearity and gene density in grass genomes. *Trends Plant Sci* **5**, 246-51.
Khan, M. Z., Zaidi, S. S.-e.-A., Amin, I., and Mansoor, S. (2019). A CRISPR Way for Fast-Forward Crop Domestication. *Trends in Plant Science*.
Kihara, H., and Ono, T. (1926). Chromosomenzahlen und systematische gruppierung der Rumex-arten. *Zeitschrift für Zellforschung und mikroskopische Anatomie* **4**, 475-481.
Kim, K. D., El Baidouri, M., Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., Libault, M., Grimwood, J., and Jackson, S. A. (2015). A Comparative Epigenomic Analysis of Polyploidy-Derived Genes in Soybean and Common Bean. *Plant physiology* **168**, 1433-1447.
Kim, M. Y., Shin, J. H., Kang, Y. J., Shim, S. R., and Lee, S.-H. (2012). Divergence of flowering genes in soybean. **37**, 857-870.
Kimber, G., and Yen, Y. (1988). Analysis of pivotal-differential evolutionary patterns. *Proceedings of the National Academy of Sciences* **85**, 9106-9108.
King, J., Bradeen, J., Bark, O., McCallum, J., and Havey, M. (1998). A low-density genetic map of onion reveals a role for tandem duplication in the evolution of an extremely large diploid genome. *Theoretical and Applied Genetics* **96**, 52-62.
Knauth, L. P., and Kennedy, M. J. (2009). The late Precambrian greening of the Earth. *Nature* **460**, 728.
Kochert, G., Stalker, H. T., Gimenes, M., Galgaro, L., Lopes, C. R., and Moore, K. (1996). RFLP and Cytogenetic Evidence on the Origin and Evolution of Allotetraploid Domesticated Peanut, Arachis hypogaea (Leguminosae). *American Journal of Botany* **83**, 1282-1291.
Koinange, E. M., Singh, S. P., and Gepts, P. (1996). Genetic control of the domestication syndrome in common bean. *Crop Science* **36**, 1037-1045.
Korneliussen, T. S., Moltke, I., Albrechtsen, A., and Nielsen, R. J. B. b. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. **14**, 289.
Kuroda, Y., Kaga, A., Tomooka, N., Yano, H., Takada, Y., Kato, S., and Vaughan, D. (2013). "QTL affecting fitness of hybrids between wild and cultivated soybeans in experimental fields."
Lackey, J. A. (1980). Chromosome Numbers in the Phaseoleae (Fabaceae: Faboideae) and their Relation to Taxonomy. *American Journal of Botany* **67**, 595.
Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W., He, W., Qin, N., and Wang, B. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature genetics* **42**, 1053.
Lande, R., and Schemske, D. W. (1985). The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39**, 24-40.
Lavin, M., Herendeen, P. S., and Wojciechowski, M. F. (2005). Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst Biol* **54**, 575-94.
Leal-Bertioli, S., Shirasawa, K., Abernathy, B., Moretzsohn, M., Chavarro, C., Clevenger, J., Ozias-Akins, P., Jackson, S., and Bertioli, D. (2015). Tetrasomic recombination is surprisingly frequent in allotetraploid Arachis. *Genetics* **199**, 1093-105.

Lee, J. S., and Verma, D. P. (1984). Structure and chromosomal arrangement of leghemoglobin genes in kidney bean suggest divergence in soybean leghemoglobin gene loci following tetraploidization. *Embo j* **3**, 2745-52.
Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in genetics* **20**, 116-122.
Lemos, B., Meiklejohn, C. D., and Hartl, D. L. (2004). Regulatory evolution across the protein interaction network. *Nature genetics* **36**, 1059.
Li, F., Kitashiba, H., Inaba, K., and Nishio, T. J. D. r. (2009). A Brassica rapa linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits. **16**, 311-323.
Li, H.-L., Wang, W., Mortimer, P. E., Li, R.-Q., Li, D.-Z., Hyde, K. D., Xu, J.-C., Soltis, D. E., and Chen, Z.-D. (2015). Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. *Scientific Reports* **5**, 14023.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60.
Li, Q. G., Zhang, L., Li, C., Dunwell, J. M., and Zhang, Y. M. (2013). Comparative genomics suggests that an ancestral polyploidy event leads to enhanced root nodule symbiosis in the Papilionoideae. *Mol Biol Evol* **30**, 2602-11.
Li, W.-H., Yang, J., and Gu, X. (2005). Expression divergence between duplicate genes. *TRENDS in Genetics* **21**, 602-607.
Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S.-s., Zuo, Q., Shi, X.-h., Li, Y.-f., Zhang, W.-k., Hu, Y., Kong, G., Hong, H.-l., Tan, B., Song, J., Liu, Z.-x., Wang, Y., Ruan, H., Yeung, C. K. L., Liu, J., Wang, H., Zhang, L.-j., Guan, R.-x., Wang, K.-j., Li, W.-b., Chen, S.-y., Chang, R.-z., Jiang, Z., Jackson, S. A., Li, R., and Qiu, L.-j. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* **32**, 1045.
Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., He, J., Xu, D., May, G., and Stacey, G. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J* **63**, 86-99.
Lin, J.-Y., Stupar, R. M., Hans, C., Hyten, D. L., and Jackson, S. A. (2010). Structural and functional divergence of a 1-Mb duplicated region in the soybean (Glycine max) genome and comparison to an orthologous region from Phaseolus vulgaris. *The Plant Cell*, tpc. 110.074229.
Lin, Y. R., Schertz, K. F., and Paterson, A. H. (1995). Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. *Genetics* **141**, 391-411.
Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., Nagamatsu, A., Arai, M., Yamada, T., Kitamura, K., Masuta, C., Harada, K., and Abe, J. (2010). The soybean stem growth habit gene Dt1 is an ortholog of Arabidopsis TERMINAL FLOWER1. *Plant Physiol* **153**, 198-210.
Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International* **2012**.
Liu, S., Luo, J., Chai, J., Ren, L., Zhou, Y., Huang, F., Liu, X., Chen, Y., Zhang, C., Tao, M., Lu, B., Zhou, W., Lin, G., Mai, C., Yuan, S., Wang, J., Li, T., Qin, Q., Feng, H., Luo, K.,

Xiao, J., Zhong, H., Zhao, R., Duan, W., Song, Z., Wang, Y., Wang, J., Zhong, L., Wang, L., Ding, Z., Du, Z., Lu, X., Gao, Y., Murphy, R. W., Liu, Y., Meyer, A., and Zhang, Y.-P. (2016). Genomic incompatibilities in the diploid and tetraploid offspring of the goldfish × common carp cross. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 1327-1332.

Lloyd, A., and Bomblies, K. (2016). Meiosis in autopolyploid and allopolyploid Arabidopsis. *Curr Opin Plant Biol* **30**, 116-22.

Logeman, B. L., Wood, L. K., Lee, J., and Thiele, D. J. (2017). Gene duplication and neo-functionalization in the evolutionary and functional divergence of the metazoan copper transporters Ctr1 and Ctr2. *Journal of Biological Chemistry* **292**, 11531-11546.

Lou, P., Wu, J., Cheng, F., Cressman, L. G., Wang, X., and McClung, C. R. (2012). Preferential retention of circadian clock genes during diploidization following whole genome triplication in Brassica rapa. *The Plant Cell*, tpc. 112.099499.

Lukens, L. N., Pires, J. C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of sequence loss and cytosine methylation within a population of newly resynthesized Brassica napus allopolyploids. *Plant physiology* **140**, 336-348.

Lutz, A. M. (1907). A PRELIMINARY NOTE ON THE CHROMOSOMES OF ŒNOTHERA LAMARCKIANA AND ONE OF ITS MUTANTS, O. GIGAS. *Science* **26**, 151-152.

Lutz, W., Sanderson, W., and Scherbov, S. (1997). Doubling of world population unlikely. *Nature* **387**, 803.

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459-73.

Ma, X. F., and Gustafson, J. P. (2005). Genome evolution of allopolyploids: a process of cytological and genetic diploidization. *Cytogenetic and Genome Research* **109**, 236-249.

Mable, B. (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society* **82**, 453-466.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5454-5459.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *J Genet* **92**, 155-61.

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., and Lysak, M. A. (2010). Fast Diploidization in Close Mesopolyploid Relatives of Arabidopsis. *The Plant Cell* **22**, 2277-2290.

Mandakova, T., and Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr Opin Plant Biol* **42**, 55-65.

Marcet-Houben, M., and Gabaldon, T. (2011). TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res* **39**, e66.

Marcker, K. A., Bojsen, K., Jensen, E. Ø., and Paludan, K. (1984). The Soybean Leghemoglobin Genes. *In* "Advances in Nitrogen Fixation Research: Proceedings of the 5th International Symposium on Nitrogen Fixation, Noordwijkerhout, The Netherlands, August 28 – September 3, 1983" (C. Veeger and W. E. Newton, eds.), pp. 573-578. Springer Netherlands, Dordrecht.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B., Steuernagel, B., Mayer, K. F., and Olsen, O. A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092.

Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* **296**, 910-913.

Mason, A. S., Nelson, M. N., Yan, G., and Cowling, W. A. (2011). Production of viable male unreduced gametes in Brassica interspecific hybrids is genotype specific and stimulated by cold temperatures. *BMC Plant Biology* **11**, 103.

Mason, A. S., and Pires, J. C. (2015). Unreduced gametes: meiotic mishap or evolutionary mechanism? *Trends in Genetics* **31**, 5-10.

Mather, K. (1940). The determination of position in crossing-over. *Journal of Genetics* **39**, 205-223.

McClintock, B. (1930). A Cytological Demonstration Of The Location Of An Interchange Between Two Non-Homologous Chromosomes Of Zea Mays. *Proceedings of the National Academy of Sciences of the United States of America* **16**, 791-796.

McClintock, B. (1993). The significance of responses of the genome to challenge.

McClintock, B., and Hill, H. E. (1931). The cytological identification of the chromosome associated with the RG linkage group in Zea mays. *Genetics* **16**, 175-190.

McLennan, H. A., Armstrong, J. M., and Kasha, K. J. (1966). Cytogenetic behaviour of alfalfa hybrids from tetraploid by diploid crosses. *Canadian Journal of Genetics and Cytology* **8**, 544-555.

McMillin, D. E., and Scandalios, J. G. (1980). Duplicated cytosolic malate dehydrogenase genes in Zea mays. *Proc Natl Acad Sci U S A* **77**, 4866-70.

Meyers, L. A., and Levin, D. A. (2007). On The Abundance Of Polyploids In Flowering Plants. *Evolution* **60**, 1198-1206.

Michael, T. P., and Jackson, S. (2013). The First 50 Plant Genomes. *The Plant Genome* **6**.

Mielczarek, M., and Szyda, J. (2016). Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics* **57**, 71-79.

Mikhailova, E. I., Naranjo, T., Shepherd, K., Wennekes-van Eden, J., Heyting, C., and De Jong, J. H. (1998). The effect of the wheat Ph1 locus on chromatin organisation and meiotic chromosome pairing analysed by genome painting. *Chromosoma* **107**, 339-350.

Milla, R., Osborne, C. P., Turcotte, M. M., and Violle, C. (2015). Plant domestication through an ecological lens. *Trends in ecology & evolution* **30**, 463-469.

Mirzaghaderi, G., and Mason, A. S. (2017). Revisiting Pivotal-Differential Genome Evolution in Wheat. *Trends Plant Sci* **22**, 674-684.

Moore, R. C., and Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Current opinion in plant biology* **8**, 122-128.

Moose, S. P., and Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant physiology* **147**, 969-977.

Morales, A. M. A. P., O'Rourke, J. A., van de Mortel, M., Scheider, K. T., Bancroft, T. J., Borém, A., Nelson, R. T., Nettleton, D., Baum, T. J., Shoemaker, R. C., Frederick, R. D., Abdelnoor, R. V., Pedley, K. F., Whitham, S. A., and Graham, M. A. (2013). Transcriptome analyses and virus induced gene silencing identify genes in the Rpp4-mediated Asian soybean rust resistance pathway. *Functional Plant Biology* **40**, 1029-1047.

Moretzsohn, M. C., Gouvea, E. G., Inglis, P. W., Leal-Bertioli, S. C., Valls, J. F., and Bertioli, D. J. (2013). A study of the relationships of cultivated peanut (Arachis hypogaea) and its most closely related wild species using intron sequences and microsatellite markers. *Ann Bot* **111**, 113-26.
Morgan, T. H., Sturtevant, A. H., Muller, H. J., and Bridges, C. B. (1922). "The Mechanism of Mendelian Heredity," Holt.
Morrell, P. L., Buckler, E. S., and Ross-Ibarra, J. J. N. R. G. (2012). Crop genomics: advances and applications. **13**, 85.
Mosse, J. (1990). Nitrogen-to-protein conversion factor for ten cereals and six legumes or oilseeds. A reappraisal of its definition and determination. Variation according to species and to seed protein content. *Journal of Agricultural and Food Chemistry* **38**, 18-24.
Muller, H. J. (1916). The Mechanism of Crossing-Over. *The American Naturalist* **50**, 193-221.
Muller, H. J. (1936). Bar duplication. *Science* **83**, 528-530.
Müntzing, A. (1936). The evolutionary significance of autopolyploidy. *Hereditas* **21**, 363-378.
Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet* **49**, 490-496.
Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., Prins, B., Reynolds, A., Chia, J.-M., and Ware, D. (2011). Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences* **108**, 3530-3535.
Nei, M., Gu, X., and Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences* **94**, 7799-7806.
Newton, n. C. F., and Pellew, n. (1929). Primula kewensis and its derivatives. *Journal of Genetics* **20**, 405-467.
Nguepjop, J. R., Tossim, H.-A., Bell, J. M., Rami, J.-F., Sharma, S., Courtois, B., Mallikarjuna, N., Sane, D., and Fonceka, D. (2016). Evidence of Genomic Exchanges between Homeologous Chromosomes in a Cross of Peanut with Newly Synthetized Allotetraploid Hybrids. *Frontiers in Plant Science* **7**.
Niederhuth, C. E., Bewick, A. J., Ji, L., Alabady, M. S., Kim, K. D., Li, Q., Rohr, N. A., Rambani, A., Burke, J. M., Udall, J. A., Egesi, C., Schmutz, J., Grimwood, J., Jackson, S. A., Springer, N. M., and Schmitz, R. J. (2016). Widespread natural variation of DNA methylation within angiosperms. *Genome Biology* **17**, 194.
Nishida, H., Yoshida, T., Kawakami, K., Fujita, M., Long, B., Akashi, Y., Laurie, D. A., and Kato, K. (2013). Structural variation in the 5′ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time. *Molecular Breeding* **31**, 27-37.
Njiokou, F., Bellec, C., Berrebi, P., Delay, B., and Jarne, P. (1993). Do self-fertilization and genetic drift promote a very low genetic variability in the allotetraploid Bulinus truncatus (Gastropoda: Planorbidae) populations? *Genetics Research* **62**, 89-100.
Nowak, M. A., Boerlijst, M. C., Cooke, J., and Smith, J. M. (1997). Evolution of genetic redundancy. *Nature* **388**, 167.
Ohno, S. (1970). "Evolution by Gene Duplication," Springer Berlin Heidelberg.
Ohno, S. (1973). Ancient Linkage Groups and Frozen Accidents. *Nature* **244**, 259.
Ohno, S., Wolf, U., and Atkin, N. B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169-187.
Okamoto, M. (1957). Asynaptic effect of chromosome V. *Wheat Inf. Serv.* **5**, 6.

Oliver, F. W. (1913). "Makers of British botany: a collection of biographies by living botanists," University press.
Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant physiology* **171**, 2294-2316.
Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290.
Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytologist* **186**, 5-17.
Parisod, C., Mhiri, C., Lim, K. Y., Clarkson, J. J., Chase, M. W., Leitch, A. R., and Grandbastien, M.-A. (2012). Differential dynamics of transposable elements during long-term diploidization of Nicotiana section Repandae (Solanaceae) allopolyploid genomes. *PLoS One* **7**, e50352.
Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A. S., Pires, J. C., Barker, G., Moore, J., Walley, P. G., Manoli, S., Batley, J., Edwards, D., Nelson, M. N., Wang, X., Paterson, A. H., King, G., Bancroft, I., Chalhoub, B., and Sharpe, A. G. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology* **15**, R77.
Pastor-Satorras, R., Smith, E., and Solé, R. V. (2003). Evolving protein interaction networks through gene duplication. *Journal of Theoretical biology* **222**, 199-210.
Paterson, A. H. (2005). Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica* **123**, 191-196.
Paterson, A. H., Chapman, B. A., Kissinger, J. C., Bowers, J. E., Feltus, F. A., and Estill, J. C. (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.
Peng, J., Ronin, Y., Fahima, T., Röder, M. S., Li, Y., Nevo, E., and Korol, A. (2003). Domestication quantitative trait loci in Triticum dicoccoides, the progenitor of wheat. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2489-2494.
Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C., and Doyle, J. J. (2005). Placing Paleopolyploidy in Relation to Taxon Divergence: A Phylogenetic Analysis in Legumes Using 39 Gene Families. *Systematic Biology* **54**, 441-454.
Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., and Marroni, F. (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular biology and evolution* **33**, 2706-2719.
Polak, R., Phillips, E. M., and Campbell, A. (2015). Legumes: Health Benefits and Culinary Approaches to Increase Intake. *Clinical diabetes : a publication of the American Diabetes Association* **33**, 198-205.
Poncet, V., Robert, T., Sarr, A., and Gepts, P. (2004). Quantitative trait locus analyses of the domestication syndrome and domestication process. *Encyclopedia of plant and crop science* **1069**, 1074.
Population Reference Bureau (2007). "2007 World population data sheet," PRB Washington, DC.

Prince, J. P., Pochard, E., and Tanksley, S. D. (1993). Construction of a molecular linkage map of pepper and a comparison of synteny with tomato. *Genome* **36**, 404-417.
Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* **40**, e11.
Purugganan, M. D., and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature* **457**, 843.
Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.
Quint, M., and Gray, W. M. (2006). Auxin signaling. *Current opinion in plant biology* **9**, 448-453.
Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and Locating Whole Genome Duplications on a Phylogeny: A Probabilistic Approach. *Molecular Biology and Evolution* **31**, 750-762.
Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* **29**, 467-501.
Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology* **14**, 3158.
Rastogi, S., and Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC evolutionary biology* **5**, 28-28.
Ratnaparkhe, M. B., Singh, R. J., and Doyle, J. J. (2011). Glycine. *In* "Wild Crop Relatives: Genomic and Breeding Resources: Legume Crops and Forages" (C. Kole, ed.), pp. 83-116. Springer Berlin Heidelberg, Berlin, Heidelberg.
Rausch, J. H., Morgan, M. T., and Husband, B. (2005). The effect of self-fertilization, inbreeding depression, and population size on autopolyploid establishment. *Evolution* **59**, 1867-1875.
RenÉE Orellana, M., López-Pujol, J., Blanché, C., and Bosch, M. (2007). "Genetic diversity in the endangered dysploid larkspur Delphinium bolosii and its close diploid relatives in the series Fissa of the Western Mediterranean area."
Renny-Byfield, S., and Wendel, J. F. (2014). Doubling down on genomes: polyploidy and crop plants. *American journal of botany* **101**, 1711-1725.
Rhoades, M. (1936). A cytogenetic study of a chromosome fragment in maize. *Genetics* **21**, 491-502.
Riley, R., Chapman, V., and Johnson, R. (2009). The incorporation of alien disease resistance in wheat by genetic interference with the regulation of meiotic chromosome synapsis. *Genetical Research* **12**, 199-219.
Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131-147.
Rodriguez, G., Muños, S., Anderson, C., Sim, S.-C., Michel, A., Causse, M., Gardener, B., Francis, D., and Knaap, E. (2011). "Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity."
Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., and Jackson, S. A. (2013). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal* **73**, 143-153.

Sakuma, S., Salomon, B., and Komatsuda, T. (2011). The domestication syndrome genes responsible for the major changes in plant form in the Triticeae crops. *Plant and cell physiology* **52**, 738-749.
Salmon, A., Ainouche, M. L., and Wendel, J. F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular ecology* **14**, 1163-1175.
Salse, J. (2016). Ancestors of modern plant crops. *Curr Opin Plant Biol* **30**, 134-42.
Salse, J., Abrouk, M., Bolot, S., Guilhot, N., Courcelle, E., Faraut, T., Waugh, R., Close, T. J., Messing, J., and Feuillet, C. (2009). Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proceedings of the National Academy of Sciences* **106**, 14908-14913.
Santos, J. L., Alfaro, D., Sanchez-Moran, E., Armstrong, S. J., Franklin, F. C. H., and Jones, G. H. (2003). Partial Diploidization of Meiosis in Autotetraploid Arabidopsis thaliana. *Genetics* **165**, 1533-1540.
Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human molecular genetics* **19**, R227-R240.
Scheid, O. M., Afsar, K., and Paszkowski, J. (2003). Formation of stable epialleles and their paramutation-like interaction in tetraploid Arabidopsis thaliana. *Nature genetics* **34**, 450.
Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution* **16**, 372-380.
Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., and Cheng, J. (2010a). Genome sequence of the palaeopolyploid soybean. *nature* **463**, 178.
Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C., and Jackson, S. A. (2010b). Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178.
Schnable, J. C. (2015). Genome Evolution in Maize: From Genomes Back to Genes. *Annual Review of Plant Biology* **66**, 329-343.
Schnable, J. C., Freeling, M., and Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome biology and evolution* **4**, 265-277.
Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* **108**, 4069-4074.
Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A.,

Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**, 1112-1115.

Schranz, M. E., Lysak, M. A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci* **11**, 535-42.

Sears, E. R. (1976). Genetic control of chromosome pairing in wheat. *Annual review of genetics* **10**, 31-51.

Sedivy, E. J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytologist* **214**, 539-553.

Seijo, G., Lavia, G. I., Fernandez, A., Krapovickas, A., Ducasse, D. A., Bertioli, D. J., and Moscone, E. A. (2007). Genomic relationships between the cultivated peanut (Arachis hypogaea, Leguminosae) and its close relatives revealed by double GISH. *Am J Bot* **94**, 1963-71.

Semon, M., and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr Opin Genet Dev* **17**, 505-12.

Seoighe, C., and Gehring, C. (2004). Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. *Trends Genet* **20**, 461-4.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *The Plant Cell* **13**, 1749-1759.

Shiu, S.-H., Shih, M.-C., and Li, W.-H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. *Plant physiology* **139**, 18-26.

Shoemaker, R. C., Polzin, K., Labate, J., Specht, J., Brummer, E. C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J. P., Kochert, G., and Boerma, H. R. (1996). Genome duplication in soybean (Glycine subgenus soja). *Genetics* **144**, 329-38.

Shoemaker, R. C., Schlueter, J., and Doyle, J. J. (2006). Paleopolyploidy and gene duplication in soybean and other legumes. *Current opinion in plant biology* **9**, 104-109.

Singh, R., and Hymowitz, T. (1988). The genomic relationship between Glycine max (L.) Merr. and G. soja Sieb. and Zucc. as revealed by pachytene chromosome analysis. *Theoretical and Applied Genetics* **76**, 705-711.

Singh, R. J. (2016). "Plant Cytogenetics, Third Edition," CRC Press.

Singh, R. J., Kim, H. H., and Hymowitz, T. (2001). Distribution of rDNA loci in the genus Glycine Willd. *Theoretical and Applied Genetics* **103**, 212-218.

Smith, B. D., and Nesbitt, M. (1995). "The emergence of agriculture," Scientific American Library New York.

Smith, B. W. (1950). Arachis hypogaea. Aerial Flower and Subterranean Fruit. *American Journal of Botany* **37**, 802-815.

Soltis, D. E., Bell, C. D., Kim, S., and Soltis, P. S. (2008). Origin and early evolution of angiosperms. *Ann N Y Acad Sci* **1133**, 3-25.

Soltis, P. S., and Soltis, D. E. (2012). "Polyploidy and genome evolution," Springer.

Song, C., Liu, S., Xiao, J., He, W., Zhou, Y., Qin, Q., Zhang, C., and Liu, Y. (2012). Polyploid organisms. *Sci China Life Sci* **55**, 301-11.
Song, F., Smith, J. F., Kimura, M. T., Morrow, A. D., Matsuyama, T., Nagase, H., and Held, W. A. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences* **102**, 3336-3341.
Song, H., Gao, H., Liu, J., Tian, P., and Nan, Z. (2017). Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in Arachis duranensis and Arachis ipaënsis orthologs. *Scientific Reports* **7**, 14853.
Song, H., Sun, J., and Yang, G. (2018). Comparative analysis of selection mode reveals different evolutionary rate and expression pattern in Arachis duranensis and Arachis ipaënsis duplicated genes. *Plant Molecular Biology* **98**, 349-361.
Song, Q., and Chen, Z. J. (2015). Epigenetic and developmental regulation in plant polyploids. *Current opinion in plant biology* **24**, 101-109.
Stebbins, G. L. (1947). Types of Polyploids: Their Classification and Significance. *In* "Advances in Genetics" (M. Demerec, ed.), Vol. 1, pp. 403-429. Academic Press.
Stebbins, G. L. (1950). "Variation and Evolution in Plants," Columbia University Press.
Stebbins, G. L. (1971). "Chromosomal evolution in higher plants," Edward Arnold Ltd., London.
Sturtevant, A. H. (1915). The behavior of the chromosomes as studied through linkage. *Zeitschrift für induktive Abstammungs- und Vererbungslehre* **13**, 234-287.
Stütz, A. M., Schlattl, A., Zichner, T., Korbel, J. O., Rausch, T., and Benes, V. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339.
Subramaniam, S., Wang, X., Freeling, M., and Pires, J. C. (2013). The fate of Arabidopsis thaliana homeologous CNSs and their motifs in the Paleohexaploid Brassica rapa. *Genome Biol Evol* **5**, 646-60.
Sybenga, J. (1996). Chromosome pairing affinity and quadrivalent formation in polyploids: do segmental allopolyploids exist? *Genome* **39**, 1176-1184.
Tan, S., Zhong, Y., Hou, H., Yang, S., and Tian, D. (2012). Variation of presence/absence genes among Arabidopsis populations. *BMC Evolutionary Biology* **12**, 86.
Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008a). Synteny and collinearity in plant genomes. *Science* **320**, 486-488.
Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**, 102.
Tang, H., Lyons, E., and Schnable, J. C. (2014). Chapter Eight - Early History of the Angiosperms. *In* "Advances in Botanical Research" (A. H. Paterson, ed.), Vol. 69, pp. 195-222. Academic Press.
Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome research*, gr. 080978.108.
Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G. C., Wang, X., Freeling, M., and Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics*, genetics. 111.137349.

Tanksley, S. D., Bernatzky, R., Lapitan, N. L., and Prince, J. P. (1988). Conservation of gene repertoire but not gene order in pepper and tomato. *Proceedings of the National Academy of Sciences* **85**, 6419-6423.
Tariq, M., and Paszkowski, J. (2004). DNA and histone methylation in plants. *TRENDS in Genetics* **20**, 244-251.
Tasdighian, S., Van Bel, M., Li, Z., Van de Peer, Y., Carretero-Paulet, L., and Maere, S. (2017). Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. *Plant Cell* **29**, 2766-2785.
TATE, J. A., SOLTIS, D. E., and SOLTIS, P. S. (2005). Polyploidy in plants. *In* "The evolution of the genome", pp. 371-426. Elsevier.
Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**, 692.
Teichmann, S. A., and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nature genetics* **36**, 492.
Teichmann, S. A., and Veitia, R. A. (2004). Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics* **167**, 2121-2125.
Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A., Wei, S., Preece, J., Geniza, M. J., Jiao, Y., Lee, Y. K., Wang, B., Mulvaney, J., Chougule, K., Elser, J., Al-Bader, N., Kumari, S., Thomason, J., Kumar, V., Bolser, D. M., Naamati, G., Tapanari, E., Fonseca, N., Huerta, L., Iqbal, H., Keays, M., Munoz-Pomer Fuentes, A., Tang, A., Fabregat, A., D'Eustachio, P., Weiser, J., Stein, L. D., Petryszak, R., Papatheodorou, I., Kersey, P. J., Lockhart, P., Taylor, C., Jaiswal, P., and Ware, D. (2018). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic acids research* **46**, D1181-D1189.
Teufel, A. I., Liu, L., and Liberles, D. A. (2016). Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization. *BMC Evol Biol* **16**, 45.
The French–Italian Public Consortium for Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463.
The Legume Phylogeny Working Group (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny The Legume Phylogeny Working Group (LPWG). *Taxon* **66**, 44-77.
The Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189.
Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome research* **16**, 934-946.
Thomas, C. V., and Waines, J. G. (1984). Fertile backcross and allotetraploid plants from crosses between tepary beans and common beans. *Journal of Heredity* **75**, 93-98.
Tian, Z., Zhao, M., She, M., Du, J., Cannon, S. B., Liu, X., Xu, X., Qi, X., Li, M.-W., Lam, H.-M., and Ma, J. (2012). Genome-Wide Characterization of Nonreference Transposons Reveals Evolutionary Propensities of Transposons in Soybean. *The Plant Cell* **24**, 4422-4436.

Tiley, G. P., Ané, C., and Burleigh, J. G. (2016). Evaluating and Characterizing Ancient Whole-Genome Duplications in Plants with Gene Count Data. *Genome Biology and Evolution* **8**, 1023-1037.
Ting, Y. C. (1966). Duplications and Meiotic Behavior of the Chromosomes in Haploid Maize (Zea mays L.). *CYTOLOGIA* **31**, 324-329.
Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2018). Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant biotechnology journal* **16**, 749-759.
Tsubokura, Y., Matsumura, H., Xu, M., Liu, B., Nakashima, H., Anai, T., Kong, F., Yuan, X., Kanamori, H., Katayose, Y., Takahashi, R., Harada, K., and Abe, J. (2013). Genetic Variation in Soybean at the Maturity Locus E4 Is Involved in Adaptation to Long Days at High Latitudes. *Agronomy* **3**, 117.
Tsuchimatsu, T., Kaiser, P., Yew, C.-L., Bachelier, J. B., and Shimizu, K. K. (2012). Recent loss of self-incompatibility by degradation of the male component in allotetraploid Arabidopsis kamchatica. *PLoS Genetics* **8**, e1002838.
Tsukaya, H. (2013). Does Ploidy Level Directly Control Cell Size? Counterevidence from Arabidopsis Genetics. *PLOS ONE* **8**, e83729.
Uhl, C. H. (1992). Polyploidy, Dysploidy, and Chromosome Pairing in Echeveria (Crassulaceae) and Its Hybrids. *American Journal of Botany* **79**, 556-566.
Valliyodan, B., Qiu, D., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., and Song, L. (2016). Landscape of genomic diversity and trait discovery in soybean. *Scientific reports* **6**, 23598.
Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. *Trends Plant Sci* **14**, 680-8.
Vanblaere, T., Szankowski, I., Schaart, J., Schouten, H., Flachowsky, H., Broggini, G. A., and Gessler, C. J. J. o. b. (2011). The development of a cisgenic apple plant. **154**, 304-311.
Vance, C. P., Graham, P. H., and Allan, D. L. (2000). Biological nitrogen fixation: Phosphorus-A critical future need? *In* "Nitrogen fixation: From molecules to crop productivity", pp. 509-514. Springer.
Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014a). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome research* **24**, 1334-1347.
Vanneste, K., Maere, S., and Van de Peer, Y. (2014b). Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130353.
Vasconcelos, E. V., de Andrade Fonsêca, A. F., Pedrosa-Harand, A., de Andrade Bortoleti, K. C., Benko-Iseppon, A. M., da Costa, A. F., and Brasileiro-Vidal, A. C. (2015). Intra- and interchromosomal rearrangements between cowpea [Vigna unguiculata (L.) Walp.] and common bean (Phaseolus vulgaris L.) revealed by BAC-FISH. *Chromosome Research* **23**, 253-266.
Veitia, R. A. (2004). Gene Dosage Balance in Cellular Pathways. *Implications for Dominance and Gene Duplicability* **168**, 569-574.
Veitia, R. A., Bottani, S., and Birchler, J. A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in Genetics* **24**, 390-397.
Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology* **23**, 148-154.

Vida, G. (1978). Genetic Diversity and Environmental Future. *Environmental Conservation* **5**, 127-132.
Vincenten, N., Kuhl, L.-M., Lam, I., Oke, A., Kerr, A. R., Hochwagen, A., Fung, J., Keeney, S., Vader, G., and Marston, A. L. (2015). The kinetochore prevents centromere-proximal crossover recombination during meiosis. *eLife* **4**, e10850.
Visarada, K. B. R. S., Meena, K., Aruna, C., Srujana, S., Saikishore, N., and Seetharama, N. (2009). Transgenic Breeding: Perspectives and Prospects. **49**, 1555-1563.
Wagner, A. (1994). Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences* **91**, 4387-4391.
Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular biology and evolution* **18**, 1283-1292.
Wang, H., Beyene, G., Zhai, J., Feng, S., Fahlgren, N., Taylor, N. J., Bart, R., Carrington, J. C., Jacobsen, S. E., and Ausin, I. (2015). CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proceedings of the National Academy of Sciences* **112**, 13729-13734.
Wang, H., Jiang, J., Chen, S., Qi, X., Fang, W., Guan, Z., Teng, N., Liao, Y., and Chen, F. (2014). Rapid genetic and epigenetic alterations under intergeneric genomic shock in newly synthesized Chrysanthemum morifolium x Leucanthemum paludosum hybrids (Asteraceae). *Genome biology and evolution* **6**, 247-259.
Wang, J., Sun, P., Li, Y., Liu, Y., Yu, J., Ma, X., Sun, S., Yang, N., Xia, R., Lei, T., Liu, X., Jiao, B., Xing, Y., Ge, W., Wang, L., Wang, Z., Song, X., Yuan, M., Guo, D., Zhang, L., Zhang, J., Jin, D., Chen, W., Pan, Y., Liu, T., Jin, L., Sun, J., Yu, J., Cheng, R., Duan, X., Shen, S., Qin, J., Zhang, M. C., Paterson, A. H., and Wang, X. (2017a). Hierarchically Aligning 10 Legume Genomes Establishes a Family-Level Genomics Platform. *Plant Physiol* **174**, 284-300.
Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., Ye, Z., Shen, C., Li, J., Zhang, L., Zhou, X., Nie, X., Li, Z., Guo, K., Ma, Y., Huang, C., Jin, S., Zhu, L., Yang, X., Min, L., Yuan, D., Zhang, Q., Lindsey, K., and Zhang, X. (2017b). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nature Genetics* **49**, 579.
Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytologist* **165**, 937-946.
Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., and Cheng, F. (2011). The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics* **43**, 1035.
Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., Guo, H., Kissinger, J. C., and Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40**, e49-e49.
Weir, B. S., and Cockerham, C. C. J. e. (1984). Estimating F-statistics for the analysis of population structure. **38**, 1358-1370.
Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *American journal of botany* **102**, 1753-1756.
Whittaker, R. H. (1962). Classification of Natural Communities. *Botanical Review* **28**, 1-239.
Williams, W. (1964). "Genetical principles and plant breeding," Blackwell Scientific Publications, Oxford.

Winge, Ø. (1917). "The Chromosomes, Their Numbers and General Importance," Carlesberg laboratoriet.
Winterfeld, G., Becher, H., Voshell, S., Hilu, K., and Röser, M. (2018). Karyotype evolution in Phalaris (Poaceae): The role of reductional dysploidy, polyploidy and chromosome alteration in a wide-spread and diverse genus. *PloS one* **13**, e0192869-e0192869.
Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. pp. 333. NATURE PUBLISHING GROUP, Great Britain.
Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 13875-13879.
Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. *PLOS Biology* **8**, e1000409.
Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The Effects of Artificial Selection on the Maize Genome. *Science* **308**, 1310-1314.
Wu, R., Gallo-Meagher, M., Littell, R. C., and Zeng, Z. B. (2001). A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. *Genetics* **159**, 869-882.
Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**, 1527-30.
Xu, C., Nadon, B. D., Kim, K. D., and Jackson, S. A. (2018). Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant, cell & environment*.
Xu, M., Yamagishi, N., Zhao, C., Takeshima, R., Kasai, M., Watanabe, S., Kanazawa, A., Yoshikawa, N., Liu, B., Yamada, T., and Abe, J. (2015). The Soybean-Specific Maturity Gene E1 Family of Floral Repressors Controls Night-Break Responses through Down-Regulation of FLOWERING LOCUS T Orthologs. *Plant Physiology* **168**, 1735-1746.
Young, N. D., Miller, J. C., and Tanksley, S. D. (1987). Rapid chromosomal assignment of mulilpie genomic clones in tomato using primary trisomics. *Nucleic acids research* **15**, 9339-9348.
Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., and Ma, H. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* **5**, 4956.
Zhang, H., Lang, Z., and Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology* **19**, 489-506.
Zhang, M., Kimatu, J. N., Xu, K., and Liu, B. (2010). DNA cytosine methylation in plant development. *Journal of Genetics and Genomics* **37**, 1-12.
Zhao, M., Zhang, B., Lisch, D., and Ma, J. (2017). Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. *The Plant Cell* **29**, 2974-2994.
Zheng, C., Chen, E., Albert, V. A., Lyons, E., and Sankoff, D. (2013). Ancient eudicot hexaploidy meets ancestral eurosid gene order. *BMC Genomics* **14**, S3.
Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PloS one* **9**, e85150-e85150.