The Under-Explored Diversity of the Soybean Genome

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Justin Emanuel Anderson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Robert M. Stupar, Adviser

December 2015

Acknowledgements

I would like to express my sincere appreciation to everyone who assisted in the projects related to this thesis. I especially want to thank my adviser Dr. Robert Stupar for his continual guidance and support over these years. His willingness to take time out of his busy schedule to have a conversation at any moment is forever appreciated. Somehow he is both the most motivated and personable scientist I know. I can not imagine a better mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my graduate research committee, Dr. Peter Morrell, Dr. Candice Hirsch, Dr. Nevin Young, and Dr. Chad Myers for their helpful comments and research suggestions. From private conversations to classroom lectures these individuals have influenced the way I now see the world as a scientist. I'd also like to thank Dr. James Anderson and Dr. Melania Figueroa for their willingness to participate last minute with challenging questions in my oral exam.

I also thank past and present members of the Stupar lab, Adrian Stec, Yer Xiong, Anna Hofstad, Mikey Kantar, Junqi Liu, Suma Sreekanta, Ben Campbell, Tom Kono, Austin Dobbels, Jean-Michel Michno, Ryan Donohue, Theresa Brenberg, Bala Pudota, Shaun Curtin, Yung-Tsi Bolon, and Fengli Fu for assistance, mentoring, and good conversation throughout my time at the University of Minnesota. I'd also like to thank the Morrell lab for including me in their weekly journal clubs and thought-provoking scientific discussions.

Thanks to fellow graduate students Ana Poets, Amy Jacobson, Prabin Bajgain, Patrick Ewing, Celeste Falcon, Tyler Teide, Hannah Swegarden, Alex Ollhoff, Yong Bao, Addie Thompson, Liana Nice, Kathryn Turner, Zenith Tandukar, Jeff Neyhart, Alex Brohammer, Rafael Della Coletta and many others for assistance in the annual plant breeding symposium, Borlaug memorial lecture, and making the APS graduate student offices enjoyable and productive places.

I thank the Minnesota Supercomputing Institute, the Microbial & Plant Genomics Institute, the University of Minnesota Graduate School, and the faculty and staff in the Department of Agronomy and Plant Genetics for their help and support in this study. I thank the United States Department of Agriculture National Needs for a fellowship funding three years of my research. I also think the United Soybean Board, National Science Foundation, and the Minnesota Soybean Research and Promotion Council for helping to fund these projects.

Lastly, I thank my family for endless support and encouragement in this process and life in general. They have taught me how to live, to love, and to serve. I hope my continued research in agriculture will benefit them and the other farmers around the world.

Dedication

I dedicate this thesis to my wife Sarah. Whether near or far she has always encouraged and motivated me. I will always be thankful.

Abstract

Genetic diversity is an important component to ongoing plant breeding. Understanding where it exists and what it comes from can influence the ability to search and detect valuable agronomic traits in the future. In this thesis I explore three avenues surrounding this topic. In the first chapter I explore the current literature and knowledge of structural variation, such as deletions and duplications, documented in the soybean germplasm. In the next chapter I describe detecting these unique genetic variants in a subset of 41 soybean breeding lines and interesting patterns shaping their frequencies. In the third chapter I explore the frequency at which these genetic variants are induced in fast neutron mutagenesis or plant genetic transformation and tissue culture. Finally, in my last chapter I explore the USDA germplasm diversity to analyze patterns of local adaptation and environmental association in *Glycine soja,* soybean's crop wild relative.

Table of Contents

List of Tables

Chapter 1:

Chapter 2:

Chapter 3: (No tables)

Chapter 4:

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 1: Structural Variation and the Soybean Genome

Genomic structural variation is an important component to genetic diversity in soybean. These large scale genomic differences are now known as the underlying genetic mechanisms for a number of important phenotypic traits. Identifying structural variants across numerous individuals at higher resolution is increasingly possible with a number of improving genome analyses platforms. Understanding where these polymorphisms occur and why some are maintained over evolutionary time has important biological and agronomic implications. This chapter elaborates on detecting, describing, and developing structural variation in the soybean genome and how to incorporate these polymorphisms in ongoing soybean research and improvement.

Introduction

Plants contain more types of genetic diversity than an "assembled genome" leads one to believe. When interested in genetics and genomics, many researchers in the recent past have focused on single nucleotide polymorphisms (SNPs). This is true in soybean, where the modern sequencing technologies and the release of the soybean genome assembly (Schmutz et al. 2010) have facilitated the detection of SNPs across numerous cultivars and accessions (Lam et al. 2010; Wu et al. 2010; Hyten et al. 2010; Chung et al. 2014; Qiu et al. 2014; Zhou et al. 2015). After implementing appropriate filtering steps, a list of thousands to millions of SNPs distributed across the genome can be developed. The convenience and predictability of this process has allowed researchers from all realms of genetics to participate in the genomics era. While hugely beneficial and impactful, this framework has generally ignored the larger scale genomic variants.

Structural variation (SV), an inexact term used to describe genomic variants generally larger than 1kb, is known to affect large portions of many plant genomes (Żmieńko et al. 2014). In addition to variation in size, SV also vary in type, including deletions, duplications, inversions, and translocations. Broadly, these polymorphisms can be described as either nucleotide content variation or genomic context rearrangements. Some recently published SV profiles in plants include apple (*Malus domestica*) (Boocock et al. 2015)*,* Arabidopsis (*Arabidopsis thaliana*) (Santuari et al. 2010; Cao et al. 2011), barley (*Hordeum vulgare L*) (Munoz-Amatriain et al. 2013), cucumber (*Cucumis* sativas) (Zhang et al. 2015), maize (*Zea mays*) (Swanson-Wagner et al. 2010; Chia et al. 2012; Hirsch et al. 2014), rice *(Oriza sativa*) (Yu et al. 2013; Schatz et al. 2014), sorghum

(*Sorghum bicolor* L.) (Zheng et al. 2011), and soybean (*Glycine* max) (Lam et al. 2010; McHale et al. 2012; Anderson et al. 2014[1]; Zhou et al. 2015) to name a few.

SV formation is not fully understood in plants but is often attributed to homologous recombination (HR) or non-homologous recombination repair mechanisms as revealed in human studies (Hastings et al. 2008). HR requires long stretches (hundreds of base pairs) of high sequence similarity. HR between regions of the genome that are not alleles, known as non-allelic homologous recombination, is one mechanism of SV formation. Repair pathways following DNA double strand breakage, such as non-homologous end joining (NHEJ), single strand annealing, or microhomology-mediated end joining (MMEJ), can also result in gene deletions. NHEJ based repair involves five or less bp of sequence homology while MMEJ involves 5-25 bp (Hastings et al. 2008). Fork stalling and template switching is a replication based mechanism proposed to result in deletions or duplications, but also potentially lead to complex SV events (Gu et al. 2008). SV can also result from other biological processes including T-DNA insertion (Kyndt et al. 2015) or transposon activity (Lisch 2013). While the frequency of each is unclear, estimates in barley (Munoz-Amatriain et al. 2013) and cucumber (Zhang et al. 2015) suggest deletions are most frequently attributable to double strand break repair mechanisms.

These large-scale genetic polymorphisms are associated with a number of fine-mapped phenotypic traits. Specific examples within the plant community include glyphosate resistance in Palmer amaranth (*Amaranthus palmeri)* (Gaines et al. 2010,

[1] Anderson *et al*. 2014 is in reference to the publication of chapter 2 of this thesis.

2011), boron tolerance and winter hardiness in barley (*Hordeum vulgare L.)* (Sutton et al. 2007; Knox et al. 2010)*,* dwarfism and flowering time in wheat *(Triticum spp.*) (Pearce et al. 2011; Díaz et al. 2012; Li et al. 2012), female gamete fitness in potato (*Solanum tuberosum*) (Iovene et al. 2013), submergence tolerance and grain size in rice (*Oriza sativa)* (Xu et al. 2006; Wang et al. 2015b)*,* reproductive morphology in cucumber (*Cucumis sativas)* (Zhang et al. 2015), and aluminum tolerance and glume formation in maize (*Zea mays)* (Wingen et al. 2012; Han et al. 2012; Maron et al. 2013). See (Żmieńko et al. 2014) for a comprehensive review in plants.

SV can modify gene expression in a number of ways (Figure 1). Whole gene deletions result in no DNA template for transcription and therefore an absence of that particular mRNA and protein. Gene duplications increase the amount of DNA template and may lead to additional transcription (higher mRNA expression) and downstream translation products (more protein). Gene dosage is therefore affected directly by the nucleotide content in these cases. Alternatively, genomic context can also affect gene dosage. For example, a rearrangement SV, such as an inversion or translocation, could move a gene from heterochromatin to euchromatin (or euchromatin to heterochromatin) resulting in transcriptional alterations due to the new genomic location.

SV events that only partially overlap a gene can have unique consequences. These events have less predictable effects on overall expression but generally result in a compromised transcript. For example, deleting a single internal exon might not affect a gene's transcription but may produce a non-functional protein missing an important domain. Deleting the first exon or promoter region could be sufficient to turn off expression of a gene entirely. An SV breakpoint overlapping coding sequences might even result in novel transcript formation from incidental alignment of exons. This is just a sampling of the types of disruptive scenarios SV can cause. In addition to modifying expression, SV such as transposon insertions (Yao et al. 2002) or inversions, also can inhibit local recombination. This can limit the ability to introgress traits from a specific locus as well as have long term evolutionary implications.

Ancient polyploidization events are an important factor in the exploration of soybean SV. Often referred to as a palaeopolyploid, the soybean genome has evidence of two relatively recent whole genome duplication events, occurring around 59 million years ago (mya) and 13 mya, respectively (Schmutz et al. 2010; Severin et al. 2011). Whole genome duplications (WGD) are often followed by a period of fractionation where rearrangements occur and copies of genes are deleted. Interestingly, even after millions of years the soybean genome has retained a large portion of its genes still present in multiple copies. Examining the context of this genetic redundancy is influential in shaping hypotheses surrounding SV.

Functional Structural Variants in Soybean

In soybean, the known examples of structural variants that influence phenotypic variation have been discovered in fine-mapping experiments. Perhaps the first such association was identified for a soybean seed coat color trait. In soybean production, it is common to find spontaneous black seed coat mutants in yellow seed coat varieties. Todd and Vodkin (1996) investigated this issue and found mutations in a cluster of chalcone synthase genes underlying this phenotype (Todd and Vodkin 1996). This family contained three genes, CHS1, CHS3, and CHS4. A duplication of CHS1 (termed dCHS1) was associated with the yellow seed coat, yellow hilum varieties. Spontaneous black seed coats in the offspring of full yellow seed varieties no longer had detectable full dCHS1 duplication. The authors also observed a reversion mutation from a yellow seed, colored hilum to a fully black seed. This spontaneous reversion was the result of a deletion in the CHS4 promoter region in seven out of ten cultivars. BAC sequencing found these gene family members occurred in multiple clusters within a confined area on chromosome 8, likely contributing to the frequent and recurring SV (Tuteja and Vodkin 2008). Gene transcription analyses were conducted to understand the effects of structural variation on gene regulation and phenotype. Unexpectedly, both spontaneous reversion mutation types, resulting in black seed coats, were associated with increased total CHS family mRNA. The duplication in dCHS1 reduced transcription in all family members and deletions in the CHS4 promoter increased transcription of all family members. Todd and Vodkin suspected this was likely due to an RNAi-like system of family wide silencing. The advent of siRNA sequencing confirmed the presence of natural RNAi targeting this gene family to explain the unique SV effects on transcription (Tuteja et al. 2009).

Recently, the Rhg1 soybean cyst nematode resistance QTL (Cook et al. 2012) was also characterized as a functional SV in the soybean genome. After many attempts at cloning this QTL, researchers discovered the resistance locus is a 31.2 kb segment encompassing five genes and arranged as a tandem duplication of varying copy numbers among accessions. The authors reported that three of the five genes within this segment are required for enhanced resistance, and haplotypes with more copies exhibited greater levels of resistance. Unlike the CHS example above, haplotypes with additional copies of the Rhg1 segment exhibited greater transcription of these genes. Furthermore, silencing any one of these three genes reduced soybean cyst nematode resistance. Conversely, simultaneous overexpression of these three genes in a susceptible line conferred enhanced resistance.

A larger screen of soybean germplasm followed these initial findings and identified a wider variety of SV states at the Rhg1 locus (Cook et al. 2014). Two types of resistant classes were identified: the three copy class, and a high copy class ranging from seven to ten copies. Phenotypic screens further confirmed a relationship between copy number and resistance level. Additionally, the three copy number genotypes had higher methylation than one-copy genotypes, as has been observed previously with duplicated genes (Rodin and Riggs 2003). The relationship between methylation and copy number variation is not yet clear in this situation.

Attempts to map genes resistant to a range of fungal and viral diseases (R-genes) frequently localize to regions of enhanced SV. One particularly active R-gene cluster in soybean is on chromosome 13 (Figure 2). Rsv1, resistance to soybean mosaic virus (SMV), is a cloned single member of this cluster of nucleotide binding site leucine rich repeat (NBS-LRR) genes (Hayes et al. 2004). The Rsv1 gene is responsible for resistance to many strains of SMV. Additional members of this R-gene cluster are also implicated in unique resistant and necrotic reactions to other SMV strains, depending on their presence or absence (Zhang et al. 2012). Furthermore, this locus exhibits higher total NBS-LRR genes in accessions with higher levels of SMV resistance.

The Rpp4 locus is another soybean disease resistance locus that exhibits a relationship between gene content variation in NBS-LRR genes and disease resistance levels. Rpp4 is one of few natural sources of resistance to Asian soybean rust. After the resistance loci was fine-mapped to a small space on chromosome 18, Meyer et al. discovered variation in the number of NBS-LRR type genes in this region (Meyer et al. 2009). Specifically, the susceptible reference type, Williams 82, had a cluster of three R-genes while the resistant cultivar had five R-genes. Within this five gene cluster, Rpp4C-4 was expressed in the resistant cultivar and the other members were nearly undetectable. Susceptible cultivars simply do not have this Rpp4C-4 gene.

These four putatively functional SV exemplify the complexity of this type of genetic polymorphism. Gene duplication, as in the case of Rhg1, might increase expression. Furthermore, a gene deletion will typically reduce/eliminate a gene's expression, as was the case for Rpp4 and Rsv1. These examples are intuitive. However, a duplication could also initiate a feedback loop, thus reducing expression of a gene or its entire family, as in the case of dCHS1 with yellow seed coat and hilum. In this case, a deletion might knockout a gene or gene family regulator and increase expression, as in the case of deleting part of CHS4 resulting in increased chalcone synthase family expression and a black seed coat.

Genome scans for SV

While the discovery of functional SV has relied on fine-mapping of specific loci, some soybean researchers have used cytogenetics to identify large SV events and genomics methods to catalog SV events genome-wide (reviewed by Chung and Singh

2008). Cytogeneticists have long been capable of documenting large chromosomal abnormalities. Microscopy based studies can detect aneuploidy, polyploidization, large inversions and rearrangements. In plants, the first cytological observations of these large scale events were performed in maize (McClintock 1931). However, such observations tend to be limited by chromosomes amenable to visualization under a microscope and rearrangements large enough for visual detection.

Genome analysis platforms are now allowing SV detection at a much finer scale in a wider range of species. These studies often use array-based techniques, next generation sequencing, or a combination of both. Comparative genome hybridization (CGH) arrays are the prevailing array-based technology for detecting SV in soybean. This technique utilizes preset probes designed to bind a specific region of DNA. Probes can be designed at adjacent locations along each chromosome, allowing for a genome-wide view of SV. With this method, two genotypes can be labeled with separate fluorescent dyes and co-hybridized to the probe array, producing a comparative fluorescent readout that indicates the relative DNA copy abundance for each genotype at each probe location. Like all array technology, this method has background technical variation that can cause low signal-to-noise ratios for a subset of probes. Furthermore, probes designed to match a reference sequence will hybridize more efficiently to that sequence than a genotype containing substitutions or small indels in the probe binding region. The number of probes developed, and therefore their spacing, limits the size of SV that can be detected (Gresham et al. 2008). Nonetheless, this technique has proved highly valuable in detecting SV in a wide assortment of species including yeast (Dunham et al. 2002), humans (Sebat et al. 2004; Iafrate et al. 2004), and soybeans (Haun et al. 2011).

The other common genome wide SV scanning platform utilizes next generation sequencing. Unlike CGH, every nucleotide of the genome could theoretically be assayed. There are four main approaches to detect SV with whole genome sequencing (Tattini et al. 2015). The most widespread technique is based on read depth variation (RDV), wherein sequence reads from a given genotype are mapped to a genome reference sequence and quantified as the number of reads mapped per genomic interval (e.g. reads per gene). RDV analysis would therefore predict that genomic regions in which few or no reads are mapped are putatively deleted, whereas regions with disproportionally high read-mapping coverage are duplicated. Generally, there is some necessary scaling to account for the non-normal distribution of read mapping.

In addition to RDVs, data from paired-end reads can also be helpful in SV detection and characterization. Read pairs can orient SV, such as detecting an inversion or determining whether a duplication is tandemly located or dispersed to a new location. Orientation and genomic location is something CGH simply cannot answer. Read pairs can also bridge deletion gaps, validating RDV detected deletions.

Split read mapping, where part of a read is masked during mapping, can also be used to detect SV and increase resolution down to a single base pair at a breakpoint. CREST (Wang et al. 2011) and BreakDancer (Chen et al. 2009) are thus far the only read pair or split read based algorithms used to detect SV in soybean (Qiu et al. 2014; Bolon et al. 2014).

The final next generation sequencing based approach incorporates *de novo* assembly, requiring much higher levels of sequence coverage. Detecting structural variation through *de novo* assembly first requires the development of scaffolds and then aligning these to a previously assembled genome for comparison. SV can then be assessed based on large scale differences between the assembled scaffolds and the reference genome (Li et al. 2014). An alternative approach begins by mapping reads to the reference genome and any unmapped reads are used to assemble scaffolds. This approach provides novel assemblies for genomic regions not found in the reference genome.

The use of next generation sequencing also has limitations (Sims et al. 2014). Plant genomes generally have a high degree of repetitive elements, making read mapping unclear or inaccurate in many genomic regions. Mistakes or misassembles in the reference genome could accidentally be interpreted as SV. Furthermore, nucleotides that do not exist in the reference genome but are present in other lines are difficult to incorporate and often ignored.

Long read sequencing technologies, now increasingly available, might alleviate some of the deficiencies of both CGH and next generation sequencing. Current long read technologies now produce single reads many kilobases in length. Single molecule sequencing, such as with PacBio, can be very helpful in reference genome assembly, particularly across highly repetitive regions (Huddleston et al. 2014). Sequencing across long genomic regions also greatly facilitates accurate SV detection (Wang et al. 2015a). While this technology is comparatively expensive and has a higher error rate in calling nucleotides (Wang et al. 2015a), techniques are being developed to account for and correct these errors, and the long read technology will undoubtedly appear in soybean SV publications in the near future.

Understanding the Limitations of a Reference Genome

The largest limiting factor for all of the aforementioned approaches is the biases associated with a single reference genome sequence. Improvements in the reference genome will help to anchor scaffolds, bridge gaps, and confirm orientation of segments. However, even a perfect reference genome assembly will not solve all of the problems. One issue is caused by genetic heterogeneity among individuals within any given soybean cultivar. Small amounts of residual intra-cultivar variation are not likely a problem for farmers, but can cause major issues in genomics. Intra-cultivar variation is primarily a consequence of the plant breeding process. Most soybean breeding strategies require only a limited number of single-seed descent generations following an initial cross, which is then followed by bulk-harvesting for seed increase and evaluation. Any remaining heterozygous regions at the time of the last single-seed descent generation will be free to segregate and differentially fix among sub-lineages of the population (which will eventually become the cultivar). The soybean reference cultivar, Williams 82, has a number of documented regions of genomic variation among such sub-lineages (Haun et al. 2011). Documentation of the cultivar release specifies that the final inbred was a combination of four separate BC_6F_3 families (Bernard and Cremeens 1988). This variation is not specific to Williams 82. Nearly all soybean cultivars are likely to have some level of intra-cultivar variation. Additionally, mutation is an ongoing process, wherein novel substitutions and SV continuously arise, resulting is slight differences between the individual plants of study and the reference genome (Ossowski et al. 2010).

Even if a perfectly inbred, mutation free, reference genome were assembled, analysis platforms based on it would still not detect all forms of variation between genotypes. For example, a gene absent in the reference but present in a different cultivar would not be detected in most current genome-wide scans. Of the soybean SV examples discussed, both Rpp4 and Rsv1 would not be detected based on strict comparisons of their source lines against the Williams 82 reference, as these genes do not exist in Williams 82. Studies of whole genome biology are attempting to address this limitation by developing species-wide genome catalogs known as a pan-genome. The idea of a pan-genome comes from the bacterial community, where scientists detected gene content variation between isolates (Tettelin et al. 2005). A few key patterns arose. The first was a subset of genes were found in all isolates, termed the "core" genome. It is presumed that all (or nearly all) individuals in the entire species has these "core" genes. Alternatively, those genes found in some but not all individuals in a species were termed the "dispensable" genes.

These pan-genome ideas of "core" and "dispensable" genome components can be applied to any species exhibiting SV, including plants (Morgante et al. 2007). It is tempting to consider the "core" genome as a list of the essential genes but this would be an over simplification. The core genome is defined by observing naturally occurring variation and therefore genes conserved across the entire species range. However, from a molecular biology perspective, essential genes are generally those that are necessary for survival in a lab or specific controlled conditions. Therefore, essential genes are likely part of the core genome but all genes in the core might not be essential for plant survival under lab conditions (Klein et al. 2012).

Genes initially classified as "dispensable" might be beneficial under certain environmental conditions (Marroni et al. 2014). A specific R-gene, for example, might be necessary to survival under a certain disease pressure, but entirely dispensable in the absence of this pressure. Genetic redundancy might also be misclassified as dispensable. An example of this occurs in Arabidopsis where following a dispersed gene duplication, divergent evolution resulted in separate lineages each carrying only one functional copy of an essential gene (Bikard et al. 2009). Since not all individuals have a copy of this gene at the same location, this essential gene is considered dispensable.

A recent publication with d*e novo* assembly of seven geographically diverse *Glycine soja* accessions is the first pan-genome analysis in soybean (Li et al. 2014). The authors estimated around 80 percent of the genome was present in all samples, making up the core genome. According to their results, thousands of genes in the pan-genome do not occur in the current *Glycine max* reference genome. Even with a limited sample size of seven genotypes, the *G. soja* pan-genome is estimated to contain 30.2 Mb more than any single individual's assembly (Li et al. 2014). Development of a complete soybean pan-genome would require *de novo* assembly of many more individuals. This level of assembly isn't currently feasible in most plant species. Instead, studies of SV often focus only on gene space. Analyzing gene space can produce a genic pan-genome, or similarly a pan-transcriptome, surmised from transcriptome data. For example, transcriptome data from a wide array of individuals and a bulked tissue type was recently used to infer a maize pan-genome (Hirsch et al. 2014).

Diversity and SV within and between *Glycine soja* and *Glycine max*

The pan-genome study of *G. soja* is one of several publications that have assayed the genomic diversity in soybean's wild relative (Table 1). Researchers are interested in *G. soja* because modern soybean lost much of its genetic diversity in the domestication and improvement process (Hyten et al. 2006). As a close relative, *G. soja* has a similar genome to soybean making it amenable to crossing or genome scans for SV.

The first SV observed between a *G. soja* and *G. max* comparison was an inversion (Ahmad et al. 1979). Since then, a number of additional inversions have been also been detected within some *G. soja* individuals (Palmer et al. 2000; Kim et al. 2010b; Qiu et al. 2014). These inversions are segregating in *G.* soja and do not represent fixed differences between the species (Palmer et al. 2000). Inversions, like those observed, naturally maintained in the wild are often associated with adaptation to clinal variation or even speciation (Kirkpatrick and Barton 2006). Inversions are inherently negative due to the deleterious meiotic consequences of unequal crossing over. In order for an inversion to be maintained and at detectable frequencies, it must include a beneficial pair of alleles (Kirkpatrick 2010). If two beneficial alleles are included in an inversion, then recombination can not separate them and the benefit of having both alleles outweighs the deleterious meiotic consequences. This concept has been discussed in the population genetics community and documented in other crop wild relatives (Fang et al. 2012). In addition to the previously discussed SV detection methods, inversions can also be discovered as regions of highly elevated linkage disequilibrium. Other factors also affect linkage disequilibrium, such as reduced recombination in soybean's large pericentromeric regions (Song et al. 2013), suggesting any putative inversions should be validated using an additional technique. As of yet, inversions detected in *G. soja* have yet to be further explored.

Translocations, though rare, have been discovered in soybean individuals (Mahama et al. 1999). Through a combination of mapping populations and cytology using fluorescence *in situ* hybridization, recent studies have characterized the chromosomes involved and approximate breakpoints of the seven known translocation events in soybean (Findley et al. 2010, 2011). These individual events were derived from a range of backgrounds including: *G. soja, Glycine gracilis* (close relative of *G. soja* and *G. max*), fast-neutron irradiated *G. max* populations, and a spontaneous translocation in a *G. max* cultivar cross. The presence of these large translocations in heterozygous individuals can produce a single chain or ring of multiple chromosomes pairing in meiosis potentially resulting in pollen or ovule sterility. One of these translocations, occurring frequently in *G. soja* accessions from northern China, might explain the occasional semi-sterility found in *G. soja* by *G. max* crosses (Findley et al. 2010). Smaller translocation events, that could be detected through the use of paired-end or *de novo* assembly and comparison, likely also occur but have yet to be explored in soybean.

Recent genome scans of *G. soja* and *G. max* have detected widespread nucleotide content SV within and between these species (Table 1). These include SV segregating in *G. max* and *G. soja* as well as those only present in *G. soja* (Li et al. 2014; Zhou et al.

2015). Many of these studies were focused on detecting SNPs and indels then followed with a scan for SV based on RDV. These resequencing studies often analyzed *G. soja* accessions, *G. max* landraces, and/or cultivars in order to detect QTL related to the domestication or improvement process. More deletions are discovered than duplications, or other types of SV, as these are most easily detected with RDV or CGH. Some disparity between these studies is linked to the number of genotypes, the depth of sequencing, and the parameters used. Based on this collection of diverse studies, in soybean elite lines, landraces, and wild relatives SV effects up to nearly ten percent of the genome and around three percent of genes. Similarly, an Arabidopsis study involving 80 lines observed RDV in around two percent of the reference genome (Cao et al. 2011). The rates of SV in maize are much higher with estimates up to 30% or more (Chia et al. 2012).

Evolving R-gene clusters

SV often occurs in genes functionally annotated as biotic stress response (McHale et al. 2012; Li et al. 2014; Anderson et al. 2014). One major family is the NBS-LRR genes, the same family responsible for most of the cloned disease resistance genes in plants (Dangl and Jones 2001; McHale et al. 2006). As demonstrated in Figure 2, these R-genes tend to occur in clusters. These clusters average nearly five NBS-LRR genes per locus (Shao et al. 2014). R-gene clustering is a pattern occurring in a wide variety of plant species studied (Michelmore and Meyers 1998) that develops from tandem and segmental duplication (Leister 2004). The locally repetitive structure of R-gene clusters can lead to additional SV through gene conversion and unequal crossing over. Rapid changes in disease resistant gene content, especially in these gene clusters, is likely an important component to evolving disease resistance (Michelmore and Meyers 1998).

NBS-LRR type R-genes generally act to directly or indirectly recognize pathogen effector proteins and trigger a defense response (Jones and Dangl 2006). This gene-for-gene interaction model between plant and pathogen results in a constantly evolving arms race (Flor 1971; Takken and Rep 2010; Ravensdale et al. 2011). Genomic studies of plant pathogens, such as *Phytophthora sojae,* have discovered SV in their avirulence genes as well (Qutob et al. 2009). In this evolutionary arms race, gene deletion and duplication appears to be an important evolutionary mechanism for both plants and pathogens. One might ask why R-gene clusters aren't constantly expanding to defend against all pathogens. In the presence of a pathogen, a specific R-gene might be essential but without this selective force the gene may be dispensable or even have a fitness costs (Tian et al. 2003; Bomblies and Weigel 2007). To explore this hypothesized fitness cost, one group created a pair of near-isogenic lines in Arabidopsis where the only variant was the gene responsible for resistance. Under non-inoculated conditions the genotype without the R-gene was found to yield 9% more (Tian et al. 2003).

Evolutionary Dynamics of SV

Gene duplications are less frequently discovered than gene deletions in genome scans but their implications are no less significant. Gene duplication has long been implicated as the route to new function (Ohno 1970). Initially, gene duplication simply results in genetic redundancy. This idea of redundancy suggests both of these paralogs are contributing to the same gene function. While potentially beneficial (Liu et al. 2008), this redundancy is often unnecessary and potentially disruptive leading to silencing, subfunctionalization, or neofunctionalization over time (Lynch and Conery 2000). Subfunctionalization is when both members of the pair accumulate degenerative mutations to a point where combined they serve the same function as the ancestral gene (Lynch and Force 2000). Neofunctionalization is when one of the duplicated pair develops into a role unrelated to its previous evolutionary function. New gene duplications arise at higher rates than nucleotide substitution rates but these new events are often under purifying selection (Katju and Bergthorsson 2013). This was evidenced in soybean by the excess of rare variants found in the frequency distribution of duplications in the soybean nested association mapping parents (SoyNAM) (Anderson et al. 2014). Studies in other species have also noted this purifying selection (Epstein et al. 2014). If the duplicated genes affect the stoichiometry in a biochemical pathway then increased dosage can be deleterious, as suggested in the gene balance hypothesis (Birchler and Veitia 2012). A more thorough discussion of the population genetic implications of gene duplication is reviewed in Katju and Bergthorsson (2013).

Deletions, unlike duplications under purifying selection, are often neutral. For example, the frequency of deletions found in the SoyNAM parents resembles a simulated neutral model (Anderson et al. 2014). This finding is a bit counter intuitive. On the surface, it suggests genes can be lost without negative consequences. While certainly not true for all genes, genetic redundancy might imply many of these copies aren’t necessary. Deletions removing pseudogenes, annotated as genes, would also be inconsequential.

Current studies of fast neutron mutagenesis lines suggest large chromosomal regions can be deleted and still result in wild type looking soybean plants (Bolon et al. 2011, 2014).

Whole Genome Duplication

The history of WGD is an important factor when discussing SV in soybean. All plant species, and many other organisms, are now believed to have undergone WGD at least once in their history. As mentioned earlier, the soybean genome has evidence of recent WGDs occurring approximately 59 mya and 13 mya (Schmutz et al. 2010). In the legume family, many individuals share the more ancient event, while the 13 mya event is exclusive to the genus *Glycine* (Shoemaker et al. 2006; Severin et al. 2011). One of these WGD events appears to be an allopolyploidization, as evidenced by two types of centromeric repeats (Gill et al. 2009). Since WGD, the soybean genome has gone through a process of unbiased fractionation, where genes present and expressed today are relatively equally derived from both ancestral genomes (Garsmeur et al. 2014). Through fractionation and diploidization the soybean genome still maintains 60 to 70 percent of its genes as paralogs (Schmutz et al. 2010; Anderson et al. 2014). One might expect these gene duplicates act as a buffer of genetic redundancy. If this were the case there should be an enrichment for SV in the duplicated subset of genes among soybean genotypes, but instead the opposite is observed (Anderson et al. 2014). SV, and especially deletions, in the SoyNAM parents were less likely to overlap WGD-derived paralogs. This same pattern is found in mammals and other vertebrates, where SV infrequently overlap preserved WGD-derived paralogs (Makino et al. 2013). Subfunctionalization, observed in many of these duplicated genes (Roulin et al. 2013), wherein both copies are now necessary, is one possible explanation for the preferential maintenance. The gene balance hypothesis also suggests SV in WGD-derived paralogs would be deleterious because they affect the stoichiometry of biochemical pathways (Birchler and Veitia 2012). Therefore, the apparent genetic redundancy found in the soybean genome does not necessarily imply full functional redundancy.

Mapping using SV as a marker

The number and dispersion of SV makes these polymorphisms also useful as markers in mapping experiments (Wang et al. 2014; Shen et al. 2015). This was recently implemented in a soybean domestication and improvement study (Zhou et al. 2015). Using resequencing data, the authors called both SNPs and RDVs in 302 phenotyped lines. With these markers, they scanned for signals of selection during domestication and improvement and conducted genome wide association study (GWAS) on a number of different phenotypes. The use of SV in GWAS was successful at detecting previously known functional structural variants. When assaying for seed coat color they detected a strong signal on chromosome 8, where variation in the chalcone synthase family is known to affect seed hilum color. When assessing soybean cyst nematode resistance, they associated major resistance with the Rhg1 SV locus on chromosome 18. Interestingly, a GWAS for plant height with SV detected four significant loci on chromosome 12, including one overlapping a strong selection signal during domestication. Incorporating SV with SNPs in association mapping can improve resolution and even aid in the detection of causative genetic variants for complex phenotypes (Stranger et al. 2007).

Inducing SV

Mutagenic irradiation, such as fast neutrons (FN) or X-rays, can induce large scale SV and unique phenotypes. Using CGH and next generation sequencing, sufficiently large SV can be easily detected in mutants. In order to develop novel soybean phenotypes for breeding and gene function applications, the soybean community has recently developed two large FN irradiated populations. The resulting unique phenotypes and detected SV for a subset of mutant lines are now publicly available on Soybase.org/mutants. FN induced SV have been associated with a number of interesting traits, including hyper nodulation (Men et al. 2002), dwarfism (Hwang et al. 2014), seed protein and oil content, and short petioles (Bolon et al. 2011, 2014). Associating detected SV with the unique phenotypes in these populations is an ongoing process. This FN mutant population database also serves as a community resource for reverse genetic studies.

The patterns of SV induced by FN mutagenesis suggest a highly malleable soybean genome (Bolon et al. 2014). Mutagenesis-induced SV can affect many more genes per locus than the SV observed in diverse germplasm scans of natural variation. Among 264 FN mutant lines assayed to date (Bolon et al. 2014), more than 40 percent of the soybean genes have been identified within at least one duplicated segment, 9 percent of the genes have been found within at least one homozygous deletion, and 19 percent have been found within at least one hemizygous deletion. Much like the SV observed in the SoyNAM, FN induced SV was enriched for genes without a retained paralog from the last WGD. These findings further enforce the WGD-derived paralogs might play essential roles in biological processes.

The advent of genome editing technologies makes targeted SV induction possible (Voytas 2013). Zinc Finger Nucleases (ZFN), TALENs, and CRISPR/Cas9 have all been demonstrated to work in soybean (Curtin et al. 2011; Haun et al. 2014; Jacobs et al. 2015; Michno et al. 2015). Using one of these technologies to simultaneously target two separate loci on one chromosome can induce SV. In Arabidopsis, simultaneous ZFNs induced large deletions and inversions (Qi et al. 2013) and in rice large deletions were induced with the CRISPR/Cas9 system (Zhou et al. 2014). The recent publication of successful gene editing in soybean (Li et al. 2015) further expands the potential for novel gene insertion, arrangement, or other SV. These new technologies will likely be instrumental tools in future studies of gene function, genome evolution, and the development of new phenotypic traits.

Conclusions

Advancements in genome scanning technologies have facilitated the accurate and precise detection of SV in plants genomes. Recognizing where SV occur and how they are maintained will continue to improve our understanding of their role in adaptation and crop improvement. Structural variation is found throughout the soybean genome, exhibiting substantial enrichment in biotic defense response genes. Future research will likely uncover additional functional SV underlying important phenotypic traits. Tapping into the currently underexplored genetic diversity in the soybean germplasm is important in the search for agronomically essential traits. Furthermore, inducing SV *de novo* through traditional or biotechnology-aided mutagenesis will be useful for generating novel phenotypic variation to enable mutation breeding and studies of gene function, genome evolution, and the limits of the soybean genome.

Table 1. Genome-wide SV genotyping studies in Soybean and *G. soja*.

SV Detection Method	**Publication**	***G. soja* /Landraces/Elite lines in SV scan**	**Coverage Depth**	**Deleted**	**Duplicated**	**Novel**	**Deleted and Duplicated**
RDV & *De novo* unmapped	(Kim, Lee, *et al.* 2010)	1/0/0	43x	32.4 Mb	-	8.3 Mb	-
De novo	(Lam *et al.* 2010)	1/0/0	80x	856 genes	-	-	-
CGH and RDV	(McHale *et al.* 2012)	0/0/4	-	672 CNV genes		-	-
RDV	(Li *et al.* 2013)	8/8/9 plus previous data	3.38x	22.3 Mb	-	-	-
RDV & *De novo* unmapped	(Chung *et al.* 2014)	6/4/6	>14x	1,737 genes	-	343 genes with plant homologues	-
De novo	(Qiu *et al.* 2014)	1/0/0	55x	-	-	10 Mb	-
BreakDancer		0/1/0	41x	8.7 Mb	-	-	-
CGH & RVD	(Anderson *et al.* 2014)	0/0/41	>2x	1,200 genes	223 genes	-	105 genes
De novo Pan-genome	(Li *et al.* 2014)	7/0/0	>83	1,179 genes	726 genes	2.3-3.9 Mb/line	73 genes
RDV	(Zhou *et al.* 2015)	62/130/110	>11x	73.6 Mb	15.14 Mb	-	-

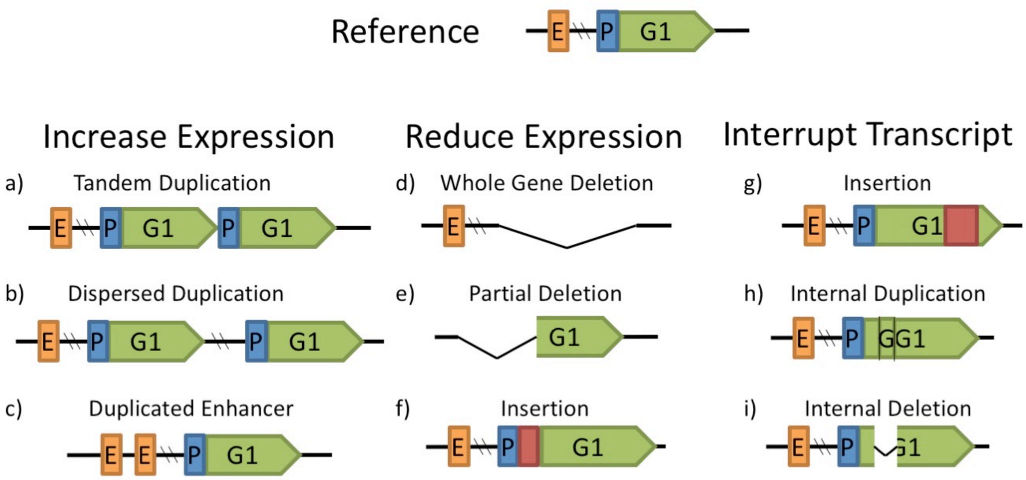


Figure 1. Potential effects of SV on gene content and transcript production. (a-c) Expression can increase as a result of tandem or dispersed whole gene duplication or duplication of an enhancer region. (d-f) Expression can decrease through whole gene deletion, partial deletion, or interruption of gene promoter region. (g-i) Internal changes can lead to interrupted genes and altered transcripts. SV detection with CGH or read depth variance are most likely to detect large scale changes (a-e) and unable to detect rearrangements or insertions (f-g). This figure is modeled after a figure in (Żmieńko *et al.* 2014).

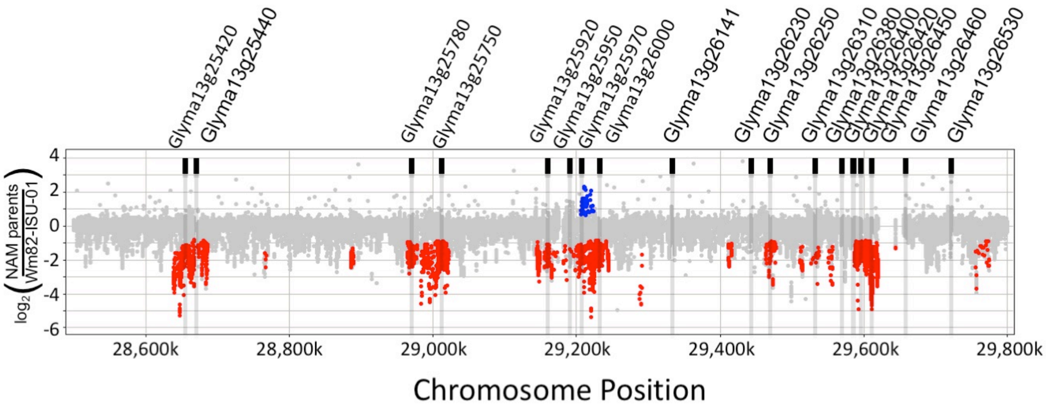


Figure 2. Structural variation in an R-gene cluster of Chromosome 13. Detected with CGH in 41 diverse soybean lines (Anderson *et al.* 2014). Plotted points are the $\log_2$ ratio of each genotype vs. the Williams82-ISU-01 reference for each probe. Colored points denote putative duplications (blue) and putative deletions (red). Labelled across the top are the location of NBS-LRR genes according to Glyma.v1.a1.1 (Schmutz *et al.* 2010). Multiple forms of disease resistance are mapped to this region including resistance to: *Phytophthora sojae* (Rps3a, Rps3b, Rps3c, and Rps8)(Gordon *et al.* 2006), *soybean mosaic virus* (Rsv1 - one of the R genes near Glyma13g25440)(Hayes *et al.* 2004; Zhang *et al.* 2012), *peanut mottle virus* (Rpv1) and *Pseudomonas syringae* (Rpg1-b) (Ashfield *et al.* 2007), and *Aphis glycines* (Soybean Aphid, Rag2)(Kim *et al.* 2010a).

Chapter 2: A Roadmap for Functional Structural Variants in the Soybean Genome[2]

Gene structural variation (SV) has recently emerged as a key genetic mechanism underlying several important phenotypic traits in crop species. We screened a panel of 41 soybean (*Glycine max*) accessions serving as parents in a soybean nested association mapping population for deletions and duplications in over 53,000 gene models. Array hybridization and whole genome resequencing methods were used as complementary technologies to identify SV in 1,528 genes, or approximately 2.8% of the soybean gene models. Though SV occurs throughout the genome, SV enrichment was noted in families of biotic defense response genes. Among accessions, SV was nearly eight-fold less frequent for gene models that have retained paralogs since the last whole genome duplication event, compared to genes that have not retained paralogs. Increases in gene copy number, similar to that described at the *Rhg1* resistance locus, account for approximately one-fourth of the genic SV events. This initial assessment of soybean SV occurrence presents a target list of genes potentially responsible for rapidly evolving and/or adaptive traits.

[2] This paper was published in G3: Genes, Genomes, and Genetics in July 2014 with the citation listed below. Supplemental methods, figures, and tables for this chapter can be found on the G3 website: http://www.g3journal.org/content/4/7/1307/suppl/DC1. Acknowledgments are listed on the next page.

Anderson, J. E., M. B. Kantar, T. Y. Kono, F. Fu, A. O. Stec, Q. Song, P. B. Cregan, J. E. Specht, B. W. Diers, S. B. Cannon, L. K. McHale, and R. M. Stupar. 2014 A roadmap for functional structural variants in the soybean genome. G3 4: 1307–1318.

ACKNOWLEDGMENT OF CO-AUTHORSHIP FOR CHAPTER 2

This was a collaborative work from a number of researchers as noted in the publication's author list. The author of this dissertation contributed to designing and performing the experiments, developing all figures and tables, CGH and resequencing data analysis, writing, and editing of this publication. MBK and TYK assisted in data analysis and developed the rSFS null hypothesis model. FF ran sequence filtering, alignment, and estimated RPKM counts. AOS ran all CGH. QA, PBC, JES, and BWD provided the SoyNAM sequence data. SBC developed the WGD-derived paralog gene list. LKM assayed the gene family enrichment. RMS assisted in experimental design, writing, and editing. All authors reviewed, commented, and approved the manuscript.

INTRODUCTION

Genome-level diversity arises from a wide spectrum of mutational events, from chromosome-level events (e.g., aneuploidy) to single nucleotide polymorphisms (SNPs). Recently there has been a surge of interest in mid-level types of polymorphism: changes smaller than chromosomal-level differences, but substantially larger than SNPs. This structural variation (SV), which is often observed as large deletions or duplications, occurs on a scale from single genes to sizeable multi-genic regions. SV segments are often referred to as copy number variation (CNV) when there is any difference in copy number across genotypes, or presence-absence variation (PAV) when some genotypes contain the segment while other genotypes are entirely devoid of the chromosomal segment.

Essentially two types of SV studies have been published in the plant research community. The first type assesses the global pattern of SV throughout the genome, using array comparative genomic hybridization (CGH) or next-generation sequencing (NGS), or a combination of these platforms. This type of study has become increasingly popular in model plant and crop species. Genome-wide SV profiles have been published recently for maize (*Zea mays*; (Swanson-Wagner *et al.* 2010; Chia *et al.* 2012), Arabidopsis (Santuari *et al.* 2010; Cao *et al.* 2011), soybean (*Glycine max*; (Lam *et al.* 2010; McHale *et al.* 2012), barley (*Hordeum vulgare* L.; (Munoz-Amatriain *et al.* 2013), and sorghum (*Sorghum bicolor* L.; (Zheng *et al.* 2011), in addition to several other species (see review by (Żmieńko *et al.* 2014)). These studies have been successful at extracting meaningful biology from the global SV patterns, but have not attempted to assess the direct impacts of an individual CNV or PAV on a particular plant phenotype.

The second type of plant SV study focuses on the association between specific CNV/PAV within genes that govern a specific trait of interest. Gene CNVs/PAVs have been associated with numerous traits of biological and agricultural importance (reviewed by (Żmieńko *et al.* 2014)). Important examples include glyphosate resistance in Palmer amaranth (*Amaranthus palmeri*; (Gaines *et al.* 2010, 2011), boron tolerance and winter hardiness in barley (Sutton *et al.* 2007; Knox *et al.* 2010), seed coat pigmentation and soybean cyst nematode resistance in soybean (Todd and Vodkin 1996; Cook *et al.* 2012), female gamete fitness in potato (*Solanum tuberosum*; (Iovene *et al.* 2013), dwarfism and flowering time in wheat (*Triticum* spp.; (Pearce *et al.* 2011; Díaz *et al.* 2012; Li *et al.* 2012), submergence tolerance in rice (*Oriza sativa*; (Xu *et al.* 2006), and aluminum tolerance and glume formation in maize (Wingen *et al.* 2012; Han *et al.* 2012; Maron *et al.* 2013). Interestingly, these studies were often initiated as map-based cloning efforts, where the mapped interval was coincident with a causative structural variant. We are not aware of any published studies where genome-wide SV profiles have been used to identify or facilitate the discovery of a candidate SV influencing a polymorphic plant trait.

Soybean is a self-pollinating species that has experienced genetic bottlenecks during domestication and modern improvement (Hyten *et al.* 2006; Li *et al.* 2013). To assess standing genomic variation in the germplasm, this study performs SV profiling on 41 soybean accessions to identify high confidence genic CNVs/PAVs. These accessions were used as parents to develop a nested association mapping (SoyNAM) population (previously described by (Stupar and Specht 2013). This panel was strategically selected for SV profiling because the SoyNAM population is now being evaluated in the Midwestern USA for several important agricultural traits. Therefore, this study serves two distinct purposes: to increase understanding of the contribution of SV to soybean genetic diversity, and to report genes impacted by CNV/PAV that might be candidate loci contributing to phenotypic variation in the SoyNAM population.

MATERIALS AND METHODS

Comparative Genomic Hybridization

'Williams 82_ISU_01' (denoted hereafter as Wm82-ISU-01) is a sub-line of the reference genome soybean (Glycine max) cultivar 'Williams 82' (Bernard and Cremeens 1988; Haun *et al.* 2011). The stock of 'Williams 82' seed containing Wm82-ISU-01 was originally obtained from Dr. Randy Shoemaker (USDA, ARS) at Iowa State University. Wm82-ISU-01 is the nearest known match to the soybean reference genome assembly version 1.0 (Schmutz *et al.* 2010; Haun *et al.* 2011), and was therefore used as the common reference for all the experiments in this study. Seeds for the 41 soybean nested association mapping (NAM) parents were obtained from Dr. James Specht at the University of Nebraska (see Supporting Information, Table S1 for a list of the NAM parents).

Seeds were planted in 4-inch pots individually containing a 50:50 mix of sterilized soil and Metro Mix. Young trifoliate leaves from 3-week-old plants were harvested and immediately frozen in liquid nitrogen. Frozen leaf tissue was powdered with a mortar and pestle in liquid nitrogen. DNA was extracted using the Qiagen Plant DNeasy Mini Kit according to the manufacturer's protocol. DNA was quantified on a NanoDrop spectrophotometer.

An updated comparative genomic hybridization (CGH) microarray designed and built by Roche NimbleGen was used that includes 1,404,208 probes. The probes were designed based on the Williams 82 reference sequence assembly version 1.0 (Schmutz *et al.* 2010). The probes, which range between 50 and 70 bp, tile the genome at a median spacing of approximately 500 bp. Labeling, hybridization, and scanning for the CGH experiments were performed as previously described (Haun *et al.* 2011; McHale *et al.* 2012). Briefly, Wm82-ISU-01 was used as the Cy5 reference in all hybridizations, while the test genotype was labeled with Cy3. The SegMt algorithm in the DEVA software was used to generate the raw data and identify segments. The program parameters were as follows: minimum segment difference = 0.1, minimum segment length (number of probes) = 2, acceptance percentile = 0.99, number of permutations = 10. Spatial correction and qspline normalization were applied.

The $\log_2$ ratio between the Cy3 and Cy5 dyes (i.e. the NAM parent genotype compared to the Wm82-ISU-01 reference) was calculated for each probe. Segments of probes were called significant if the mean of the $\log_2$ ratio was above the upper threshold or below the lower threshold for that given genotype comparison. The lower threshold for each comparison was set at three standard deviations below the $\log_2$ ratio mean. The upper threshold for each comparison was set at two standard deviations above the $\log_2$

ratio mean. Thresholds were separately calculated for each genotype comparison. A custom Perl script was used to process the DEVA generated segments for each genotype and recognize segments beyond these thresholds. (The determination of thresholds is explained in greater detail in the File S1 and in Table S2). Significant segments found below or above their respective thresholds were initially classified as 'DownCNV' and 'UpCNV,' respectively. Collectively, these segments were referred to as 'CGH Segment CNV.'

Observations of the initial analysis revealed that while DEVA segmental clustering was successful at merging and detecting large CNV regions it often did not detect smaller (e.g. gene sized) CNV and had occasionally merged such features into non-significant segments. This motivated a second methodology for calling significant CNV using individual CGH probes. To do this, the probes within or overlapping genic space were averaged to get a probe based $\log_2$ ratio score for each gene. Genes that did not overlap with any probes were assigned the overlapping DEVA segment average or the average score of the nearest two probes. Genes exhibiting average probe $\log_2$ values above or below the significance thresholds (as defined in the previous paragraph) were classified as 'DownCNV' and 'UpCNV,' respectively. Collectively, these genes were referred to as 'CGH Probe CNV.' Visual displays of the CGH data were generated using Spotfire DecisionSite software.

Whole Genome Sequence Data

DNA isolation and whole genome sequencing for each of the 41 NAM parent lines was conducted at the USDA facility in Beltsville, MD. Approximately 40 freezedried seeds of each NAM genotype was ground to a powder with a steel ball using a Retsch MM400 Mixer Mill at 30 hz for two minutes. DNA was extracted from the ground seed tissue using the Qiagen DNEasy Plant DNA isolation kit. The DNA was fragmentased for 25 min at 37°C using the NEB Next dsDNA fragmentase (NEB, Beverly, Mass) and run on an agarose gel for size selection to obtain fragments in the 400-600 bp range. An 'A' overhang was added to the ends of the fragments. The end repaired DNA libraries were ligated with the Illumina paired-end sequencing multiplex adapters (Illumina, San Diego, CA). Illumina Paired End libraries were sequenced for 150 bp on an Illumina HiSeq 2000. The reference line Wm82-ISU-01 was sequenced on an Illumina HiSeq 2000 at the University of Minnesota, using a Paired End library and 100 bp reads. Before aligning to the reference, the raw reads were cleaned using minimum base quality score Q30. Following this cleaning, the NAM 'hub' parent, IA3023 (which was mated to each of the other 40 NAM parents), was sequenced to a depth of 31x. Read depth was variable among the remaining 40 NAM parent lines, ranging from approximately 2x to 8x coverage (Table S1). Wm82-ISU-01 was sequenced to a depth of approximately 13x. The cleaned reads were mapped to the reference genome using BWA MEM (Li and Durbin 2009b). The alignments were then cleaned by removing reads: 1) that failed vendor quality check; 2) that were PCR or optical duplicates; 3) that are not properly paired; and 4), that mapped to multiple positions.

The number of sequence reads uniquely mapped between the start and stop codons of each gene were counted. Genes that had zero reads across all genotypes (including Wm82-ISU-01) were removed from further analyses. To control for scaling issues, genes that exhibited zero reads in Wm82-ISU-01 and more than one read in at least one NAM parent line were analyzed in parallel. Additionally, genes exhibiting reads in Wm82-ISU-01 and zero reads in at least one NAM parent line were flagged as potential DownCNV and also analyzed separately. RPKM (defined as Reads mapped Per Kilobase per Million mapped reads) was calculated across genes and genotypes to standardize the variable genotype coverage and gene size. For each gene, the $\log_2$ ratio of the NAM parent RPKM divided by the Wm82-ISU-01 RPKM was calculated. Using the same methods as described above for CGH analysis, genes with $\log_2$ ratios two standard deviations above the mean were considered potential UpCNV and $\log_2$ ratios below three standard deviations from the mean were considered potential DownCNV for each genotype. Collectively, these genes were referred to as 'Sequence CNV.'

Cross-validation of CGH and sequence data to find significant genes

As described above, CGH and re-sequencing analyses provided three lists of putative structural variants associated with genomic regions: 'CGH Segment CNV,' 'CGH Probe CNV,' and 'Sequence CNV.' A subset of genes were identified from these lists for downstream analysis, including: (1) Genes found within the 'CGH Segment CNVs'; (2) Genes found on both the 'CGH Probe CNV' and 'Sequence CNV' lists (Figure S1). For this subset of genes, the sequence-based $\log_2$ RPKM ratio values were plotted against the CGH-based $\log_2$ ratios for all 41 NAM parent genotypes. Structural variants were considered cross-validated among the two platforms when the 41 genotypes clearly split into two or more clusters or collectively clustered beyond stated thresholds.

See Figure S2 for a methodological flow chart from data type to CNV cross-validated calls.

The UpCNV and DownCNV classifications were subdivided into more specific categories based on the cross-validation analyses. Estimates of gene copy number per genotype were used as the criterion for classifying each gene into one of six categories, which were designated as follows. (1) DownCNV/PAV: One copy in Wm82-ISU-01, zero copies in at least one NAM parent, no more than one copy among all 41 NAM parents; (2) UpPAV: Zero copies in Wm82-ISU-01, a single group of one or more copies in at least one NAM parent (Wm82-ISU-01 had few or no reads mapped to these genes while at least one NAM parent exhibited numerous such reads skewing the RPKM based estimates); (3) UpPAV & UpCNV: Zero copies in Wm82-ISU-01, multiple groups of one or more copies among the NAM parents; (4) UpCNV & DownCNV: One copy in Wm82-ISU-01, zero copies in at least one NAM parent, more than one copy in at least one NAM parent; (5) UpCNV: One copy in Wm82-ISU-01, more than one copy in at least one NAM parent; (6) Multi-Allelic UpCNV: One copy in Wm82-ISU-01, multiple groups of one or more copies among the NAM parents.

Enrichment Analyses

Individual gene categories were analyzed for enrichment of protein domains. Protein domains were predicted for the longest open reading frame of each *Glycine max* v1.1 gene model (http://www.phytozome.net/soybean) by Pfam, with gathering thresholds defining prediction cutoffs (Finn *et al.* 2010). For simplicity of presentation, significant results from the 11 PFAM models for Leucine Rich Repeat domain containing proteins were described as a single PFAM clan (PFAM clan ID: CL00022). Enrichment of predicted protein domains in each gene list was determined by a hypergeometric distribution with adjustment for multiple hypotheses testing by resampling methods implemented with FuncAssociate 2.0 using 10,000 simulations (Berriz *et al.* 2009).

Paralogs retained from the most recent soybean WGD were identified using QUOTA-ALIGN (Tang *et al.* 2011), using parameters "--merge --self --min_size=5 --quota=1:1" - to merge local synteny blocks, in a genome self-comparison, with a minimum block-size of 5 genes, to find the paralogs from the most recent duplication. This analysis was run using the predicted amino acid sequences of the *Glycine max* v1.1 gene models (Gmax_v1.1_189_peptide.fa; http://www.phytozome.net/soybean) for cv. Williams 82. Initial anchor points (paralog candidates for QUOTA-ALIGN) were calculated using blastp from the NCBI blast+ package. Genes that were called CNV and contained a homoeologous pair were noted and frequency calculated. Statistical analysis was conducted using the R Statistical software package (R Development Core Team 2011).

Simulations

Coalescent simulations (Hudson 2002) were used to compare the site frequency spectrum (SFS) for CNV to those expected under a neutral history in a panmictic population. Hudson's MakeSamples (ms) generates infinite-sites (Kimura 1969) genetic data under a neutral coalescent process, with specified population-scaled per-locus mutation rates, recombination rates, and migration rates. For CNV, however, a peer-acceptable mutational model does not exist for estimating the per-locus mutation rate.

There are, however, map-based recombination rates (Du *et al.* 2012) and population-scaled mutation rate estimates based on DNA resequencing data (Hyten *et al.* 2006).

Previously published estimates of the population per-bp mutation rate (θ_W) (Hyten *et al.* 2006) were used to estimate the effective population of soybeans. This parameter is related to the effective population size by the equation $\theta_W = 4N_e\mu$, where N_e is the effective population size, and μ is the per-bp mutation rate. We solved this equation for Ne, using $\mu \approx 7\times10^{-9}$ per-bp, as previously estimated (Ossowski *et al.* 2010), which yielded an effective population size estimate of 29, 642.

A locus was defined as a single CGH segment, which was experimentally found to be approximately 14kb on average. The loci were treated as independent and non-overlapping in the simulations. The observed number of CNV events was used to estimate the mutation rate parameter (theta) for the simulations. An estimate of the map-based recombination rate (Du *et al.* 2012) was used for the recombination rate. The cM/Mb recombination rate estimate was converted into a per-locus rate, with a locus consisting of one CGH segment. The per-locus recombination rate was then multiplied by our estimate of the N_e, yielding a population-scaled recombination parameter of 21.54.

Site Frequency Spectra

Development of a reference-based site frequency spectrum (rSFS) required clustering of adjacent CNV and estimating frequency in the population. Development of an Up rSFS used all genes in the UpCNV and Multi Allelic UpCNV sub classes while the Down rSFS only used the DownCNV/PAV subclass due to the higher confidence and the simplification to a biallelic model. Assuming nearby genic CNV were the result of a single CNV event and using “CGH Segment CNV” calls as a guide, adjacent cross validated CNV from the mentioned classes were collapsed into segments. Frequency estimates for individual segments required at least one gene in a segment in a genotype to exceed thresholds for both CGH and resequencing-based SV calls. See Tables S3 and S4 for specific gene segmentation.

A neutral reference-based site frequency spectrum was generated from the simulation output from MS (Hudson 2002). An SFS in the typical fashion could not be constructed, since the CGH data are heavily ascertained. That is, the CGH data are an all-by-one comparison rather than a pairwise comparison, as MS creates. Therefore, the first chromosome in the MS output was designated as the “reference” and differences were counted from the reference chromosome. Since ‘0’ denotes the ancestral state (presence) and ‘1’ denotes the derived state (absence), every site that had a ‘1’ in the reference was discarded. The result is that the SFS is built from sites where Wm82 has the “ancestral” state, and the other genotypes have the “derived” state. The neutral simulations and empirical CNV distribution were then compared for only the DownCNV and UpCNV classes. The CNV distributions were based on segments rather than individual genes by analyzing only segments with cross-validated genes within the DownCNV/PAV and UpCNV classes. Segment CNV distributions for the rSFS more properly reflect the mutational model in which CNV likely originate as segments and not gene by gene.

RESULTS

Genome-wide patterns of structural variation among the soybean NAM parent lines

The soybean NAM parents, which include a diverse set of individuals from breeding programs and international introductions, represent a relatively wide sampling of 41 different accessions within maturity groups II-V (Table S1). Initial analyses of deletions and duplications among these soybean NAM parent lines were conducted using a 1.4 million feature comparative genomic hybridization (CGH) tiling microarray platform. Comparative hybridizations were performed between each of the 41 lines (labeled with Cy3 dye) and the reference genome genotype 'Wm82-ISU-01' (labeled with Cy5 dye, referred to as 'Wm82' henceforth). Figure 1 is an overlay of the 41 CGH comparisons across the twenty chromosomes. Values plotted in red denote genomic segments that are putatively absent in at least one of the 41 NAM parent lines; these were classified as "CGH Down segments." Blue peaks denote genomic segments that either (a) exhibit copy number gains relative to Wm82 in at least one NAM parent line, or (b) are present as a single copy in at least one NAM parent line but are absent in Wm82; these were classified as "CGH Up segments." The CGH analysis identified changes in hybridization intensity contributing to an average of 282 Down and 34 Up segments per NAM parent line relative to Wm82.

Resequencing data on the 41 NAM parent lines and Wm82 was used to cross-validate the CGH segment data and better estimate the deletion and duplication rates associated with predicted gene models (gene models were based on annotation version 1.1). Reads mapped Per Kilobase per Million mapped reads (RPKM) values were used to estimate gene copy number from resequencing data. Estimates of gene copy number based on RPKM ratios were compared to those based on the CGH data. Genes with similar copy number estimates in both CGH and Illumina resequencing across genotypes were considered "cross-validated" and were thence included in the downstream analyses. The cross-validated gene set included 339 gene models exclusively associated with Up regions, 1100 gene models exclusively associated with Down regions, and 89 gene models associated with both Up and Down regions among various NAM parents.

Cross-validation between the CGH and resequencing data also identified regions of presumed heterogeneity within some of the 41 NAM parent lines. DNA from approximately 40 plants was bulk-isolated from each line for the resequencing platform, whereas a single individual plant was sampled for the CGH platform. Therefore, some SV genes which reside in regions of intra-cultivar heterogeneity could be identified as exhibiting SV on one platform while matching Wm82 on the other platform. Examples of such heterogeneity are shown in Figure S3, both for a series of genes linked in a PAV region (A) and genes exhibiting UpCNV (B). Heterogeneity among samples was particularly problematic for lines 4J105-3-4, LD02-4485, LG03-3191 and LG04-4717 (the parents to NAM populations 03, 12, 25 and 26, respectively).

A database was developed to make all the processed CGH and RPKM data publicly available (http://stuparlabcnv.cfans.umn.edu:8080/). Data for all loci are reported along with scatterplots that compare the CGH and RPKM values.

Sub-classification of SV profiles and identification of potential gain-of-function variants

To better describe the range of structural variation observed across the NAM parental lines, each of the cross-validated genes were placed into one of six categories (Figure 2; Table 1). Down segments, as shown in Figure 1, are referred to as either Down copy number variants (DownCNV) or Down present-absent variants (DownPAV). The simplest interpretation of the CGH data is that many Down structural variants are DownPAV, given that the CGH platform was purposefully designed with probes that have one unique match (one copy) in the 'Williams 82' reference genome sequence. Therefore, significant Down segments were not distinguished into subclasses, and instead were classified as a single 'DownCNV/PAV' category.

Cross-validated Up genes were sorted into the five remaining categories (Figure 2). Any Up genes that were also identified as Down in at least one other NAM parent line were placed into a class designated 'UpCNV & DownCNV'. The remaining Up genes were sorted according to their inferred presence-absence status in Wm82-ISU-01 and their mode of copy number distribution among the genotypes (bimodal or polymodal) (Figure 2 and Table 1; see Materials and Methods section for additional details on the classification criteria). Table S5 gives the full list of gene models that were placed into each of the six categories.

Approximately 72% of the 1528 cross-validated genes were placed in the DownCNV/PAV class (Table 1). An additional 205 genes were placed into other 'content variant' classes, which are interpreted as being present in some genotypes while absent in others (Figure 2 and Table 1).

There were four categories in our classification system that included genes that are duplicated in some genotypes, but are not duplicated in Wm82 or other lines. These categories (which all include 'UpCNV' in the name; see Figure 2) encompass a total of 328 genes. The five genes located within the soybean cyst nematode resistance QTL *Rhg1* represent a clear example of this type of variation. The variants of the resistant *Rhg1* phenotype have been attributed to the tandem duplication (up to 10-fold) of a 31-kb interval that includes these genes on chromosome 18 (Cook *et al.* 2012). One copy of this interval, as found in the reference genome of 'Williams 82', is associated with the SCN susceptibility locus (*rhg1*). An allele with three copies of the 31-kb interval has intermediate resistance (*Rhg1-a*), whereas an allele with ten copies confers the highest known level of resistance (*Rhg1-b*) (Cook *et al.* 2012). Our cross-validated analysis confirmed the presence of at least these three different classes of *Rhg1* copy number among the soybean NAM parents (Figure 3).

A small number of gene models exhibited a SV profile similar to *Rhg1*, in which multiple (≥3) copy number classes were observed among the NAM parents. One such example is Glyma13g04670 (named Glyma.13g068800 in the annotation version Wm82.a2.v1), which is embedded within an approximately 10-15 kb segment on chromosome 13 that exhibits at least four different copy number levels (Figure 4). The Glyma13g04670 gene has been uncharacterized in soybean, but it has been annotated as a Cytochrome P450 with similarity to *Arabidopsis CYP82C4* (Murgia *et al.* 2011).

Sequence reads that map to the approximate boundaries of the duplicated ~10-15-kb segment were individually analyzed in genotypes with either one copy or multiple copies of Glyma13g04670. Genotypes with multiple copies of Glyma13g04670 showed reads mapping to chromosome position 4.971 Mb at one end, then position 4.958 Mb at the other end (Figure S4). This indicates that the increased copy number of Glyma13g04670 in these genotypes is at least partially caused by a tandem duplication of a ~14-kb interval spanning from position 4.958 Mb to 4.971 Mb on chromosome 13.

Population Analysis and SV enrichment patterns

The lists of genes associated with the six cross-validated structural variation categories were investigated for enrichment within Pfam predicted protein classes (Finn *et al.* 2010). This analysis indicated an enrichment in the protein domains characteristically encoded by resistance genes (*R*-genes), including Leucine Rich Repeat (LRR), Nucleotide Binding (NB), and Toll Interleukin Receptor (TIR) protein domains (Table 2; (Kruijt *et al.* 2005; McHale *et al.* 2006)). In contrast, enrichment of other protein domains in genes unrelated to disease resistance was not consistently evident among the examined SV categories (Table 2).

The next set of analyses focused on the duplicated nature of the soybean genome. Soybean is often referred to as a paleopolyploid, as it retains remnants of whole-genome duplications (WGDs) that occurred approximately 13 Mya (in the *Glycine* genus), and approximately 59 Mya (shortly after early diversifications in the legume family) (Schmutz *et al.* 2010). An even older genome triplication is also apparent in comparisons of some regions of the soybean genome (Severin *et al.* 2011). Soybean retained a large proportion of duplicate genes from the most recent WGD – published estimates ranging from ~43-68% of genes retained (Schmutz *et al.* 2010; Severin *et al.* 2011). In our analysis, approximately 60% (32,464/53,833) of the soybean gene models from annotation version 1.1 have retained a syntenic paralog, the vast majority of which are presumed to be derived from the most recent WGD (Table S6). Genes with retained syntenic paralogs were substantially underrepresented among the gene content variants list (Table 1). Among all categories, SVs were found in only 0.75% (244/32,464) of genes with retained syntenic paralogs, whereas CNVs were found in 6.0% (1,284/21,459) of the genes that have not retained a syntenic paralog. This represented an eight-fold difference between the two groups of genes. However, this difference was not as severe for the quantitative UpCNV categories (e.g. UpCNV was identified in ~0.22% of genes with syntenic paralogs and ~0.57% in genes without syntenic paralogs; Table 1).

For genic SV segments, the number of NAM parent lines that exhibited differences compared to Wm82 was analyzed to look for evidence of deviations from a neutral evolution null hypothesis. This analysis included the 117 Up segments (mean of 13580 bp; median of 3182 bp) and 547 Down segments (mean of 14958 bp; median of 2775 bp) that overlap with at least one gene identified as CNV/PAV. The frequency of lines showing significant differences compared to Wm82 was calculated for each of these segments. Experimental observations were used as parameters of approximate segment size for simulation of a neutral model under the coalescent. As shown in Figure S5, Down segments closely reflected the frequency spectrum of the simulated neutral model.

For Up segments the frequency spectrum is skewed toward an excess of singleton variants; i.e., those observed only in only one NAM parent line (Figure S5).

DISCUSSION

In this study, we identified genic SV events in the genomes of 41 genetically diverse soybean lines. The observed SV data confirmed major trends previously observed in a smaller analysis of just four soybean accessions. Those trends included an enrichment of SV genes arranged in tandemly-duplicated blocks, and an association of SV variation with genes contributing to biotic stress responses (McHale *et al.* 2012). Moreover, with the larger dataset obtained in this study, a much more detailed analysis was possible, which provided more definitive evidence for the broader patterns that influence soybean genome diversity, particularly regarding duplicated genes and the distribution of SV frequencies.

Paleopolyploidy is a major defining feature of the soybean genome, which experienced two whole genome duplication events approximately 59 and 13 million years ago (Schmutz *et al.* 2010). A majority of soybean genes are present in at least two copies, and a large percentage of these genes have retained duplicates since the most recent genome doubling event. It has been suggested that this feature makes soybean a difficult system for use in functional genomics, as gene redundancy will buffer the effects of mutagenesis on plant phenotypes. Given the large number of duplicate genes present in soybean, one might expect that the retained duplicates might frequently acquire SV because the loss or functional alteration of duplicate genes may not have a deleterious outcome due to its “backup” copy and, of course, could provide new opportunities for phenotypic plasticity. However, in this study, we found that genes that have retained paralogs from the most recent WGD event are underrepresented for associations with SV. This trend was most striking in the PAV events. These findings are likely due in part to enrichment of SV in hyper-variable regions, where WGD-derived duplicates may be lost (or not detected) due to local gene-cluster expansions and contractions. However, the low rate of SV in regions with retained WGD-derived paralogs also suggests that retention of these duplicate genes may be biologically significant, either due to diversification of biological functions (e.g. neofunctionalization or subfunctionalization; Roulin *et al.* 2013) or for maintaining proper stoichiometry within regulatory networks (in concordance with the gene balance hypothesis (Birchler and Veitia 2012)). These results coincide with patterns found in mammals and other vertebrates, where preserved WGD-derived paralogs often exhibit low rates of SV across the populations (Makino *et al.* 2013). Taken together, the global trend of SV data in soybean suggests that the “core” set of soybean genes maintained throughout the domesticated germplasm includes a high percentage of ancient homoeologous/duplicate genes that have been retained since the most recent polyploidization event. However, experimental biases may also contribute to this observation, as both the CGH platform design and resequencing data analyses require unique sequence tracts to detect a specific gene model; such unique sequences are less abundant among duplicated genes.

A preliminary assessment of SV frequency patterns was conducted by comparing those patterns with a simulated neutral model site frequency for Up and Down genomic segments located within genic regions. The data indicated that UpCNV regions are enriched for rare variants. This stands in contrast to what has been observed at the *Rhg1* locus, where additional copies of a 31-kb segment increases tolerance to soybean cyst nematode (Cook *et al.* 2012). Clearly, haplotypes with increased copies of *Rhg1* are actively being selected by breeding programs. However, there is growing evidence that gene copy number gains may oftentimes be detrimental to fitness (Katju and Bergthorsson 2013).

This poses an interesting question: Can SV profiles be used to predict which copy number changes might provide an adaptive advantage? One could argue that an SV profile of *Rhg1* (Figure 3) may have facilitated the cloning of this locus, as the striking copy number increase for these genes may have immediately established them as candidates located within the mapped interval. Based on the assumption that an increase in copy number confers phenotypic novelty due to altered transcription state, it is reasonable to expect that genes with copy number increases found in multiple genotypes (and at multiple different copy number levels) may be more likely to confer adaptive (and selected) traits, as with *Rhg1* (Cook *et al.* 2012). One such gene from the current study is the cytochrome P450 gene Glyma13g04670, which exhibited a full a spectrum of copy number states (up to approximately ten copies) among the 41 soybean accessions. This is a particularly interesting candidate because there are several published examples of P450 genes acting in biotic and abiotic stress response, as well as herbicide tolerance pathways (Schuler and Werck-Reichhart 2003; Saika *et al.* 2014).

The potential adaptive effect of SV remains largely unexplored. While the association of SV genes in defense gene clusters has long been known (Michelmore and Meyers 1998), there is mounting evidence that copy number gains in specific genes can have tremendous effects on abiotic stress tolerance. Previous studies in barley and maize have specifically identified copy number gains and presence-absence variants that provide enhanced tolerance to stressed soil conditions, such as boron and aluminum toxicity (Sutton *et al.* 2007; Maron *et al.* 2013). Discovery of such loci will become increasingly relevant for the soybean community as crop production expands into poorer soils, or as soils continue to accumulate heavy metals and other chemicals after years of intensive agriculture. The parental CNV and PAV data obtained in these 41 NAM parents will be increasingly useful when the progeny of the NAM parent matings are evaluated for agronomic phenotypes (to be released in May 2015) and potentially stress-related phenotypes in the future.

Table 1. The number of gene models identified within six structural variation categories. The first two rows indicate the definition of each category based on the observed presence and copy number differences between Wm82-ISU-01 and at least one of the 41 NAM parent lines. The second two rows indicate the number of genes exhibiting each category among all genes and the subset of genes that maintain a syntenic paralog.

	Gene models evaluated	**DownCNV/ DownPAV**	**UpPAV**	**UpCNV & UpPAV**	**UpCNV & DownCNV / PAV**	**UpCNV**	**Multi-Allelic UpCNV**
Wm82-ISU-01 Copy Number	1	1	0	0	1	1	1
NAM Parent Copy Number	-	0	1 or >1	>1 and (1 or >>1)	>1 and 0	>1	>1 and >>1
Genes with syntenic paralog	32464	149	4	1	10	71	9
Genes without syntenic paralog	21369	951	96	15	79	122	21
Total genes assessed	53833	1100	100	16	89	193	30

Table 2. Gene models with specific Pfam domains are enriched for associations with SV. The number of gene models expected to be associated with SV is shown, compared to the number of gene models observed to be associated with SV for each category.

		Total in soybean genome	DownCNV/ PAV		UpPAV		UpCNV & UpPAV		UpCNV & DownCNV/ PAV		UpCNV		Multi-allelic UpCNV	
Pfam ID	**Description**		**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**
CL0022	Leucine Rich Repeat	1110	168**	23	7	2	3	0	17**	2	22**	4	6	1
PF07714	Protein tyrosine kinase	786	38*	16	0	1	1	0	4	1	3	3	3	0
PF08263	Leucine rich repeat N-terminal domain	550	74**	11	1	1	0	0	9**	1	10	2	3	0
PF00931	NB-ARC domain	454	112**	9	6	1	6**	0	13**	1	9	2	2	0
PF01582	Toll-Interleukin receptor	196	30**	4	3	0	3	0	2	0	0	1	0	0
PF14368	Probable lipid transfer	104	14**	2	0	0	0	0	0	0	0	0	0	0
PF12819	Carbohydrate-binding protein of the ER	95	14**	2	0	0	0	0	2	0	1	0	0	0
PF14111	Domain of unknown function (DUF4283)	82	10*	2	0	0	0	0	2	0	1	0	0	0
PF13947	Wall-associated receptor kinase galacturonan-binding	71	10*	1	0	0	0	0	1	0	2	0	0	0
PF14380	Wall-associated receptor kinase C-terminal	33	10**	1	0	0	0	0	0	0	0	0	0	0
PF05686	Glycosyl transferase family 90	20	7**	0	0	0	0	0	0	0	0	0	0	0
PF05018	Domain of unknown function (DUF667)	7	5**	0	0	0	0	0	0	0	0	0	0	0
PF00499	NADH-ubiquinone/plastoquinone oxidoreductase chain 6	2	0	0	2*	0	0	0	0	0	0	0	0	0

Significance of enrichment was determined by Fisher's exact test with a resampling approach to correct for multiple hypotheses as implemented by the FuncAssociate 2.0 (Berriz *et al.* 2009) program using 10,000 simulations (*P < 0.01, **P < 0.001). Only Pfam domains significantly enriched (P < 0.01) in at least one SV category were listed.

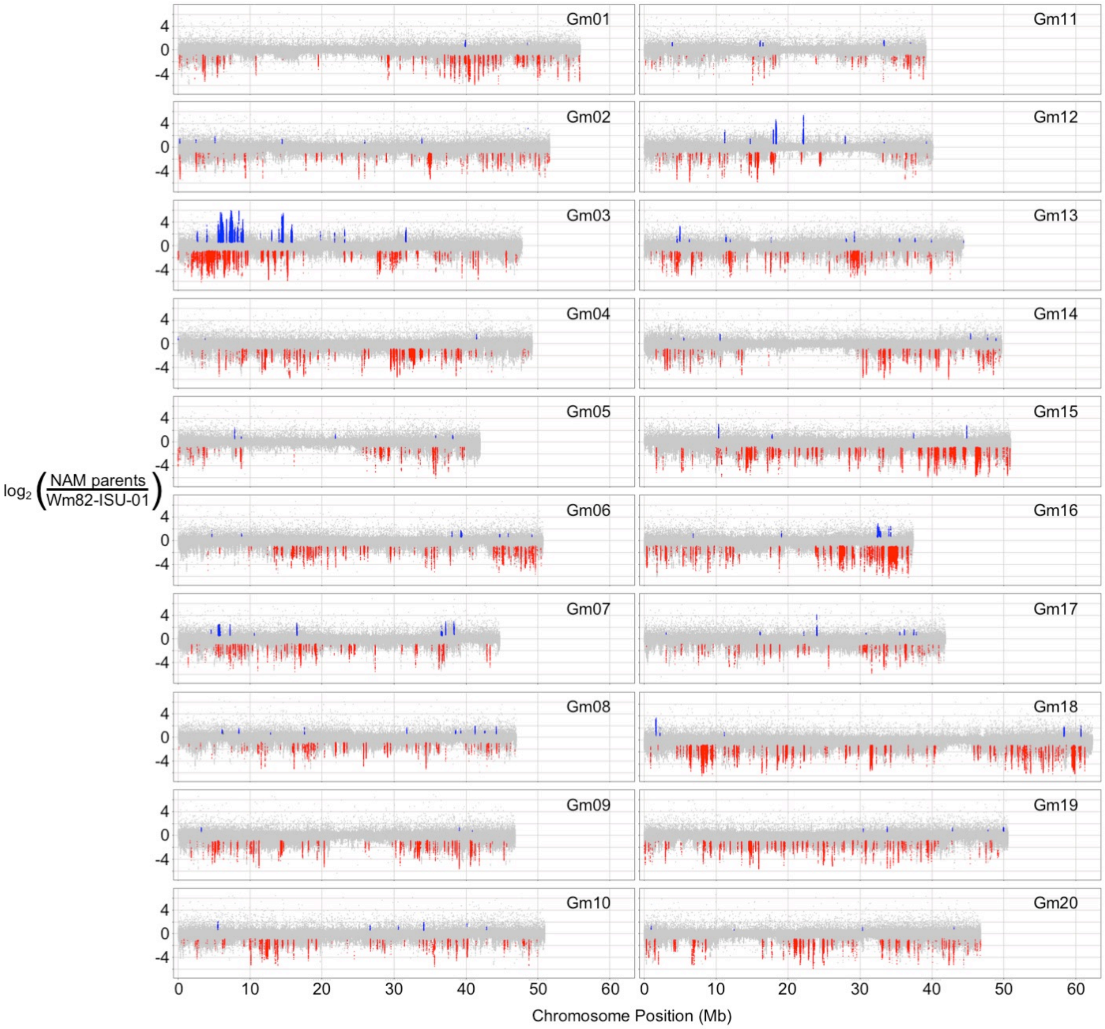


Figure 1. Genome-wide view of copy number variation found in the soybean NAM parents. Data points are the log2 ratio of each genotype versus the Williams82-ISU-01 reference for each probe. Colored spots denote probes within segments that exceed threshold; blue for UpCNV, red for DownCNV.

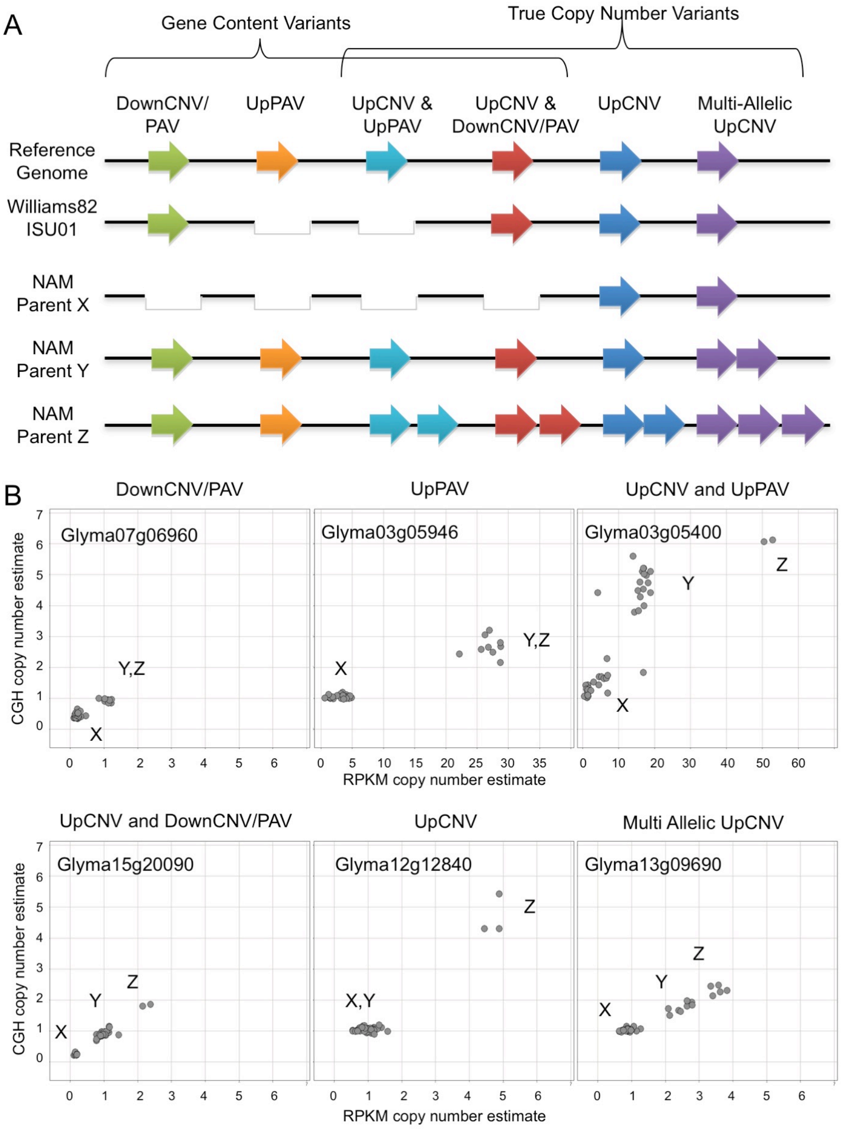


Figure 2. Classification system for CNVs that were associated with gene models. A) Presence-absence and copy number status for a hypothetical gene in each of the six classes. Genes are found in one of three states: single copy, absent (white gap), or multiple copies (two or more arrows). B) Gene representatives for each of the six classes showing allelic clusters. Each gene shows one data point for each of the 41 genotypes. The estimated copy number from sequence depth and CGH are respectively shown on the X and Y axes.

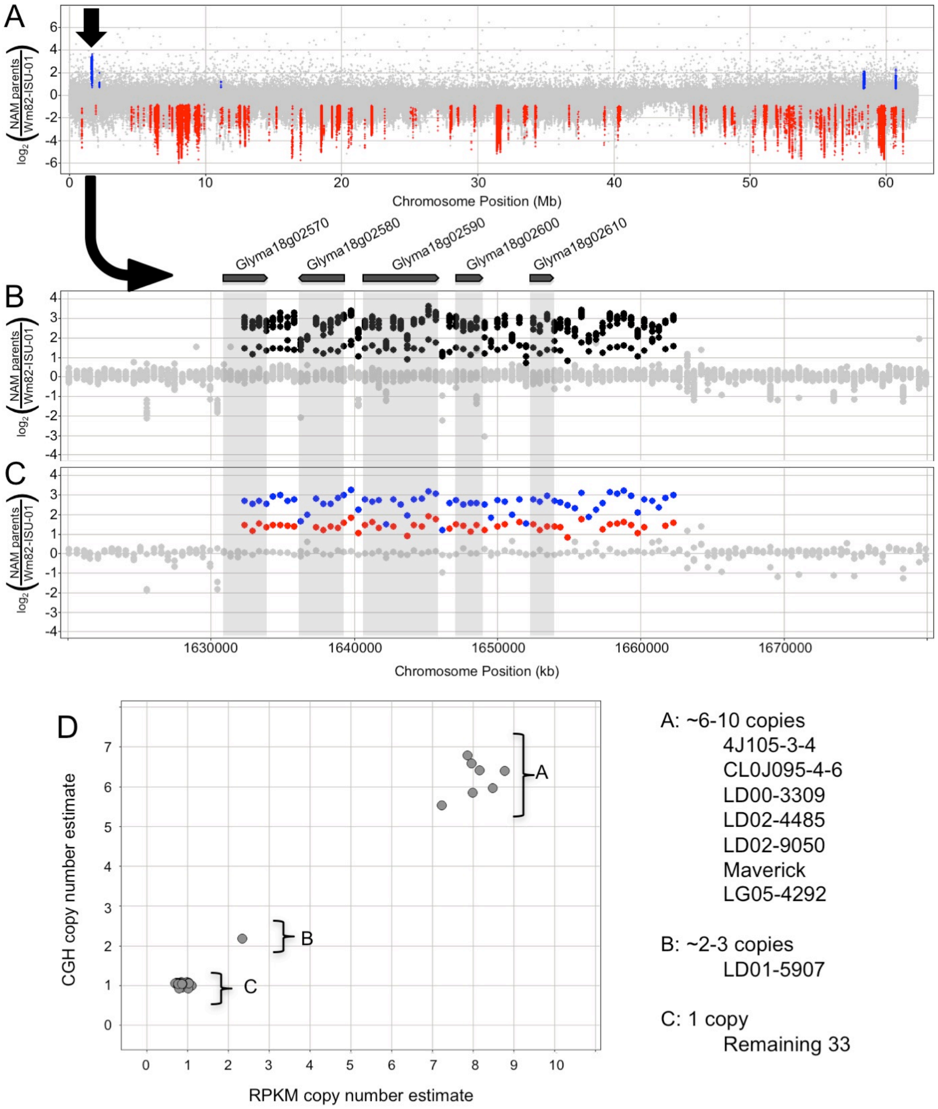


Figure 3. Copy number variation at the soybean cyst nematode locus *Rhg1*. A) The copy number variant (arrow) is clearly visible from a full view of the Chromosome 18 CGH results, overlaying data from all 41 genotypes. B) The view from (A) is zoomed in on the 31-kb UpCNV segment that overlaps five gene models (Cook *et al.* 2012). C) Viewing only one genotype from each allele class confirms a clear separation between three different copy number states. D) Cross-validation of the CNV for Glyma18g02590 using both CGH (y-axis) and sequence depth (x-axis) analyses.

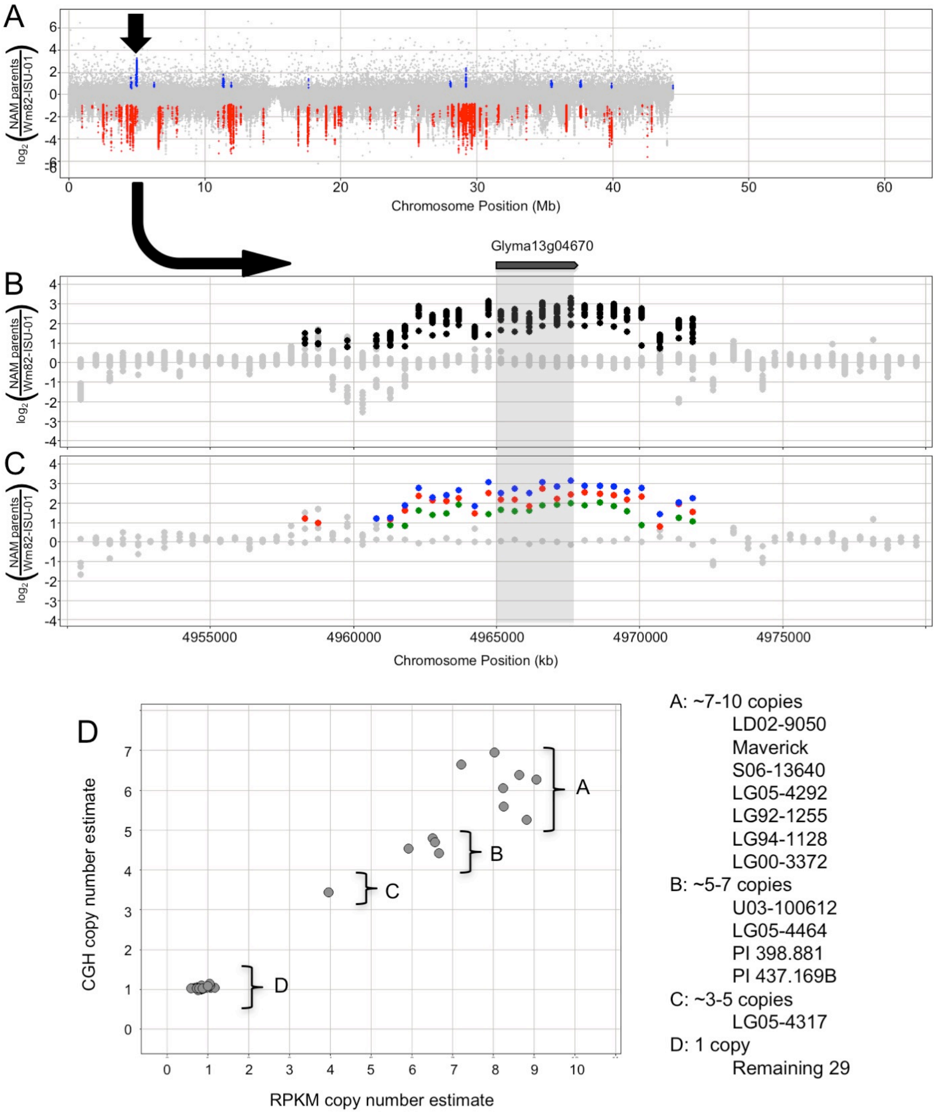


Figure 4. Copy number variation at Glyma13g04670. A) The copy number variant (arrow) is visible from a full view of the Chromosome 13 CGH results, overlaying data from all 41 genotypes. B) The view from (A) is zoomed in on the approximately 10-kb UpCNV segment that overlaps with Glyma13g04670, revealing multiple CNV classes. C) Viewing one genotype from each predicted class confirms distinct copy number states. D) Cross-validation of the CNV for Glyma13g04670 using both CGH (y-axis) and sequence depth (x-axis) analyses, revealing at least four copy number classes.

Chapter 3: Comparison of genomic variation associated with cultivars, mutagenized, and transgenic soybean plants[3]

The safety of mutagenized and genetically transformed plants remains a subject of scrutiny, despite scant information about the genomic variation induced by these technologies. In this study, genomic structural variation (e.g. large deletions and duplications) and single nucleotide polymorphism rates were assessed among a sub-sample of soybean cultivars, fast neutron-derived mutants, and genetically transformed plants. On average, transgenic plants exhibited genic structural variants one order of magnitude less than fast neutron mutants and two orders of magnitude less than rates observed between cultivars. Structural variants in transgenic plants, while rare, occurred at the transgene locus and on different chromosomes, and exhibited sequence microhomology at the repair junctions. The single nucleotide substitution rates were modest in both fast neutron and transformed plants, exhibiting fewer than 100 substitutions genome-wide, while inter-cultivar comparisons identified over one-million substitutions. Overall, these patterns provide a fresh perspective on the genomic variation associated with induced genetic variation.

[3] This chapter is the result of collaborative research. Co-authors include: Jean-Michel Michno, Thomas J. Y. Kono, Adrian O. Stec, Benjamin W. Campbell, Shaun J. Curtin, and Robert M. Stupar. The author of this dissertation contributed to designing and performing the experiments, data analysis, development of figures and tables, writing, and editing. Supplemental data can be found in Appendix 1. JM ran NGS filtering, aligning, calling SNPs, figure 6, and the tissue culture pathway. TJYK aided in NGS pipeline and read depth estimates. AOC ran all CGH and visual inspection. BWC and SJC developed the transgene constructs. RMS helped conceive and design the study and supervised analysis. All authors reviewed, commented, and approved the manuscript.

INTRODUCTION

Plant breeders use standing variation found in elite and diverse lines as the primary source for cultivar development and trait improvement. In some cases, traits of interest cannot be found within the current germplasm. Mutagenesis or genetic transformation provides avenues to introduce these traits. Standard mutagenesis methods alter DNA sequences at random loci throughout the genome in an attempt to generate novel trait variation. Genetic transformation, alternatively, attempts to insert one or few transgenes to confer a novel trait. Genetic transformation in crop species requires plant tissue culture methods. Somaclonal variation, an unintended consequence of plant tissue culture, encompasses genetic and epigenetic changes that can result in heritable phenotypic traits (Neelakandan and Wang 2012). Because such unintended changes may theoretically compromise the safety of transgenic plants (Latham *et al.* 2006), it is important to understand the coupled effects of genetic transformation and tissue culture (Schnell *et al.* 2015) and how these compare to standing and other types of induced variation.

Naturally occurring somaclonal variation is a well-established source of novel phenotypes in many vegetatively propagated fruits and vegetables, where they are commonly known as ‘sports’. Somaclonal variation induced through tissue culture, first observed in sugarcane (*Saccharum*) (Heinz and Mee 1971), has been reported in many other plant species (Neelakandan and Wang 2012). Desirable agronomic traits and released cultivars have even been derived from this type of induced variation (Jain 2001). The molecular underpinnings of somaclonal variation can include DNA sequence changes, chromosome rearrangements, aneuploidy, activation of transposable elements, and epigenetic restructuring (Neelakandan and Wang 2012). Genome-wide single nucleotide changes resulting from tissue culture have been recently observed using high-throughput sequencing in Arabidopsis (Jiang *et al.* 2011) and rice (Miyao *et al.* 2012; Zhang *et al.* 2014; Endo *et al.* 2014). These studies suggest tissue culture increases the single nucleotide mutation rate and may activate transposons (Sabot *et al.* 2011).

The insertion of a transgene is also known to create localized or dispersed genomic changes. Recent studies found that transformation can result in DNA inserted at multiple loci, multiple transgenes per locus, fragmented T-DNA, and chromosome rearrangements (Nacry *et al.* 1998; Muskens *et al.* 2000; Svitashev and Somers 2002; Clark and Krysan 2010), though such complex events are rare and discarded rather than commercialized. According to a study in Arabidopsis, transgene insertion is generally random across chromosomes, in both genic and non-genic sequences, and frequently associated with a deletion ranging from 11 to 100 bp in size (Forsbach *et al.* 2003). For soybean (*Glycine max)*, *Agrobacterium* based transformation methods occasionally result in multiple insertion sites, tandem insertions, and integration of plasmid backbone sequences (Olhoft *et al.* 2004). Recently, resequencing methods have been used to accurately localize and resolve transgene insertions (Kovalic *et al.* 2012; Kanizay *et al.* 2015). While advanced technologies have helped detect local and dispersed effects of tissue culture and transformation, limitations still exist due to sequencing errors, genetic heterogeneity of plant accessions, and reference bias (Sims *et al.* 2014).

Separating the changes induced by transformation from existing genetic variation can be a challenge (Ladics *et al.* 2015). Plant genomes can vary dramatically between cultivars. A large portion of this variation occurs as genomic structural variants (SV), such as large deletions and duplications (Żmieńko *et al.* 2014). These SV are associated with a number of biological and agriculturally important traits (Żmieńko *et al.* 2014). Previous studies in soybean have used array-based comparative genomic hybridization (CGH) or resequencing approaches to observe levels of standing SV among accessions (McHale *et al.* 2012; Anderson *et al.* 2014), or SV induced through fast neutron (FN) mutagenesis (Bolon *et al.* 2014). However, no comparable studies have addressed the incidence of tissue culture and transformation on rates of genome-wide SV in soybean.

This study investigates five transgenic (T_1 generation) soybean plants derived from standard *Agrobacterium*-mediated transformation. SV in these five lines was assessed by CGH and two of these lines were resequenced to ascertain the frequency of nucleotide substitutions. These data allow for comparisons of genomic variation in transgenic plants to the genomic variation observed in mutagenized and standing accessions. These analyses provide new insight towards understanding somaclonal variation, the effects of transgene insertion, the inheritance of SV, and the genomic consequences of developing mutant and transgenic stocks as compared to standing variation already present in soybean germplasm.

RESULTS

Genome-wide structural variation

A CGH tiling microarray with 1.4 million features was used to detect genome-wide SV in three classes of germplasm. The first class consisted of five transgenic plants each derived from a unique transformation event. Each transgenic plant contains a different transgene (Table S1), transformed using *Agrobacterium*. A range of different transgene types are represented, including a green fluorescence protein (GFP) transgene, an RNAi hairpin, a zinc-finger nuclease (ZFN), a transcription activator-like effector nuclease (TALEN), and an mPing-Pong transposon. Genotyping was done on the T_1 generation. Genome-wide CGH screens for deletions and duplications revealed single, unique novel SV in four of the five genotypes. These consisted of three deletions and one duplication (Table S1). The plant WPT_312-5-126 (ZFN transgene) did not exhibit any SV.

The second class, sampling FN induced variation, consisted of a sub-set of 35 lines from a larger mutant population developed in the genotype 'M92-220' (Bolon *et al.* 2014). These lines exhibited no obvious mutant phenotypes, and were thus referred to as "no-phenotype". The final class, representing inter-cultivar variation, came from a previous study of genic SV (Anderson *et al.* 2014), and consists of 41 parental lines from a soybean Nested Association Mapping (SoyNAM) population.

All three datasets (transgenic, FN, and inter-cultivar) were designed to detect SV in each individual genotype as compared to an appropriate reference (Supplementary Table 2). The transgenic plants were compared to the transformation parent line ('Bert' for four of the plants and ‘Williams 82’ for one plant; see Table S2), the FN plants compared to the mutagenesis parent line (‘M92-220’), and the SoyNAM parents were compared to the reference genotype ‘Williams 82’. The Methods section includes analysis details and information on how extant heterogeneity within the background cultivars was addressed.

As shown in Figure 1, CGH results varied by chromosome and by class. In this figure each black dot represents a single probe’s $\log_2$ ratio score. Clusters of dots above or below zero are putative duplications or deletions, respectively. Inter-cultivar variation, shown as the comparison of SoyNAM parent LD02-9050 to Williams 82 (Fig. 1a), occurs frequently and on nearly every chromosome. The amount of inter-cultivar variation is strikingly high when compared to a FN or transgenic plant (Fig. 1b and Fig. 1c, respectively). SV observed in FN or transformed plants generally occurred a limited number of times, on very few chromosomes, and was easier to detect.

Within the inter-cultivar class, duplications overlapped with 45 to 124 genes per cultivar comparison, while deletions overlapped with 156 to 362 genes per cultivar comparison (Fig. 2). The FN class had a lower median genic SV per line (Table S3) but was highly variable, as duplications overlapped with 0 to 1568 genes and deletions overlapped with 0 to 236 genes per line. The average size of the SV in the FN lines was over 500,000 bp, substantially larger than those observed by the inter-cultivar class whose average was less than 15,000 bp (Table S3). Of the four SV events in the transgenic plants, only two affected gene space. This included one deletion in plant WPT_389-2-2, which affected four genes on chromosome 11 (Fig. 3) and a duplication that encompassed two genes on chromosome 13 in plant WPT_301-3-13 (Fig. 4). Overall, the average number of genes affected by CGH-detectable SV in transgenic plants was estimated to be one order of magnitude less than induced by FNs and two orders less than observed among soybean varieties.

Validation of SV in the transgenic plants

The four incidences of SV detected with CGH in the transgenic plants were confirmed using PCR. Two SV events overlapped with genes, including a 125,228 bp deletion on chromosome 11 in WPT_389-2-2 (Fig. 3) and a 6,869 bp duplication on chromosome 13 in WPT_301-3-13 (Fig. 4). The two non-genic deletions were 23,406 bp in size on chromosome 1 in WPT_384-1-1 (Fig. S1) and 7,854 bp on chromosome 19 in WPT_391-1-6 (Fig. S2). Sequence data from all four SV junctions showed evidence of microhomology-mediated DNA repair (Fig. 3c, Fig. 4c, and Figs. S1c and S2d).

Screening a subset of these SV by PCR confirmed they were not intra-cultivar variation in the ‘Bert’ or ‘Williams 82’ backgrounds, as is known to exist at some loci (Haun *et al.* 2011) (Fig. S3), or derived from contamination or outcrossing from other lines (Fig. S4). The deletions on chromosome 1 and chromosome 11 were stably inherited in T_1 siblings and T_2 offspring (Figs. S1 and S5), indicating these events were both present in their respective T_0 generations. The deletion on chromosome 19 was homozygous and therefore present in the T_0 generation assuming SV is induced on a single chromosome and then becomes a homozygous deletion through genetic segregation. These data indicate these SV were derived *de novo*. The duplication on chromosome 13, however, is not found in any individual other than the T_1 transgenic genotype, WPT_301-3-13. The offspring ($T_{1:2}$), siblings (T_1), and parent (T_0) of this individual were all tested and showed no evidence of the duplication on chromosome 13 (Fig. S6). This evidence suggests the duplication arose in a post transformation generation and may not be directly attributable to the transformation process.

Transgene insertion sites

Transgenic lines were analyzed for number of transgene insertions and location of transgene(s). Southern blots of siblings or parents of WPT_301-3-13, WPT_312-5-126, and WPT_389-2-2 each showed evidence for single locus integration (Fig. S7). Thermal Asymmetric Interlaced PCR (TAIL-PCR) mapped the single insertion sites in WPT_389-2-2, WPT_384-1-1, and WPT_301-3-13. Resequencing data were also used to localize the T-DNA insertion site in WPT_389-2-2 and WPT_391-1-6. Transgene results are summarized in Table S1. Transgenes were all found to occur on different chromosomes than the aforementioned SV (Table S1). Transgene insertion and repair was observed to coincide with microhomology between the genome and the left border (Fig. 5 and Fig. S8).

According to resequencing data, transgene insertions in WPT_389-2-2 and WPT_391-1-6 induced adjacent deletions too small for CGH detection. These were the only two transgenic lines resequenced. As outlined in Figure 5a, the transgene (an mPing-Pong transposon construct) in WPT_389-2-2 induced two deletions and a 6-bp insertion of filler sequence in the T-DNA integration process. This transgene integration and associated mutations occurred in the promoter region and 5’UTR of Glyma13g33960. The WPT_389-2-2 T-DNA and adjacent mutations were homozygous in this T_1 line. The resequencing data aligned to the transgene found nine read-pairs that spanned the mPing-Pong portion of the construct (Fig. S9a) suggesting one of the homologous chromosomes has a transgene where this mPing-Pong portion was deleted or jumped out (Fig. S9b), as has been demonstrated with this element (Hancock *et al.* 2011). Had this transposon reintegrated in the genome, the methodology used for transgene mapping should have detected it. The transgene insertion in the other resequenced transgenic plant, WPT_391-1-6, also induced an adjacent ~1,200 bp deletion (Fig. S10).

Genome-wide single nucleotide substitutions

Resequencing data were used to assess the frequency of nucleotide substitutions within the inter-cultivar, FN, and transgenic classes. Based on earlier studies, it has been established that pairwise comparisons of soybean cultivars typically identify over one-million single base substitutions (Lam *et al.* 2010; Zhou *et al.* 2015). We tested our substitution identification pipeline by resequencing cultivars 'Archer' and 'Noir 1'. These data corroborated earlier studies, as 'Archer' and 'Noir 1' respectively exhibited 1,110,325 and 1,904,061 homozygous substitutions compared to the soybean reference genome 'Williams 82'.

Resequencing data were then used to asses the frequency of nucleotide substitutions in ten previously sequenced FN lines and the FN parent 'M92-220' (Bolon *et al.* 2014). These ten lines were not the same "no-phenotype" FN lines used for the CGH analysis, however were considered an acceptable alternative as they had SV frequencies similar to the no-phenotype lines (Table S4). Substitutions were detected and filtered so only those homozygous and novel to one line were included. This filtering method was based on previous mutation accumulation studies (Ossowski *et al.* 2010; Jiang *et al.* 2011; Belfield *et al.* 2012). The FN mutagenized lines had on the order of tens of unique homozygous substitutions per line (Table S5), with the highest line exhibiting 73 substitutions. However, most of these substitutions may be attributed to spontaneous processes (Ossowski *et al.* 2010) rather than the FN treatment, as the nonmutagenized 'M92-220' control also exhibited 41 unique substitutions relative to the ten FN lines. As shown in Figure 6a, substitutions in the FN lines were distributed across many more chromosomes than SV.

The two resequenced transgenic plants also showed few homozygous and novel substitutions (Table S5). The number of novel homozygous base-pair substitutions per line were as follows: two in line WPT_391-1-6, 18 in line WPT_389-2-2, one in the first 'Bert' control plant, and two in the second control 'Bert' plant. The location of the substitutions in the transgenic plants appeared unrelated to the location of the transgene insertion or the induced SV (Fig. 6b) and did not occur in coding regions (Table S5).

DISCUSSION

In this study, we observed the rates of SV and single nucleotide substitutions in transgenic and FN lines to explore a genetic component of the unintended consequences of these breeding practices. The primary safety concern relating to these previously unassessed genomic changes is that novel genetic variants might disrupt genes or pathways leading to an unforeseen harmful byproduct (Latham *et al.* 2006). For simplicity in this comparative analysis, we assume each individual gene deleted or duplicated results in the same, albeit low, new risk of a harmful byproduct. We therefore focused on the number of new mutations rather than a specific risk associated with any given mutation or mutagen. Differences in the number of induced genomic variants, attributed to an increased mutation rate, serves as the proxy for the amount of risk in unintended consequences of these breeding practices.

Under these assumptions, the level of SV across these three classes has interesting implications. The SV observed in the inter-cultivar comparison is widespread throughout the genome, repeatedly found in multiple lines, and frequently encompass only a single gene. This diversity has developed through ongoing spontaneous mutation over countless generations. Each of the genetic variants seen in this class would not represent a new risk to consumers, as any associated byproducts likely already exist in the current marketplace. The genetic variation currently segregating in these elite lines is only a subset of the total genetic diversity found in *Glycine max* or the wild progenitor *Glycine soja* (Lam *et al.* 2010; Zhou *et al.* 2015). Genetic variation arising spontaneously, or introgressed from diverse lines into elite cultivars, is a process by which even cultivars developed through traditional breeding methodology unintentionally introduce novel variants to the marketplace.

The SV observed in the no-phenotype FN lines contrasts with the patterns of SV in the inter-cultivar class. SV induced through FN mutagenesis are oftentimes large and highly variable from line to line in terms of the number of genes affected. This outcome is unexpected, as multigene deletions and duplications are anticipated to cause noticeable phenotypic changes.

The transgenic class had so few SV that direct comparisons are difficult. The events observed through CGH are moderate in size and impact a combined total of only six genes among the five plants. It is unclear if this corresponds to a single generation increase in the SV mutation rate as the spontaneous SV mutation rate in soybeans is not known. Working under the aforementioned assumption that each gene deleted or duplicated is a safety risk concludes the transgenic lines analyzed are of lower risk than many of the FN lines. While these transformation-induced events seem inconsequential when compared to those induced through FNs or found as standing variation, the finding that tissue culture and/or transformation is associated with *de novo* formation of novel SV is noteworthy.

Transgene insertion can be a locally disruptive event. The discovery of locally induced deletions, the addition of filler sequence, and microhomology between the left border and the insertion site, corroborate previous patterns of T-DNA insertion in Arabidopsis (Forsbach *et al.* 2003). The ~1kb deletions at transgene insertion sites in both of the resequenced lines are larger than the deletions found in Arabidopsis, but are not sufficient to confirm a pattern of large deletion-associated transgene insertion. The repeated presence of short sequence homology at the T-DNA insertion sites and the breakpoints of the four SV observed at non-transgene loci in these plants, implies the microhomology-mediated end joining pathway (McVey and Lee 2008) may be involved in DNA repair of these events.

The use of FN mutagenesis or tissue culture/transformation has been previously reported to result in a single generation increase in single nucleotide substitutions (Jiang *et al.* 2011; Belfield *et al.* 2012; Miyao *et al.* 2012; Zhang *et al.* 2014; Endo *et al.* 2014). A single nucleotide substitution disrupting a coding or regulatory region could similarly have an assumed safety risk associated with a novel byproduct. The FN lines and transgenic plants in this study accumulated a similar number of unique homozygous substitutions to a subset of previously published results. For example, a FN mutagenesis study in Arabidopsis detected between 5 and 18 novel homozygous substitutions per M3 line (Belfield *et al.* 2012) and a similar study of Arabidopsis tissue culture reported between 9 and 65 novel homozygous substitutions per R1 (equivalent to T1) line (Jiang *et al.* 2011). Unexpectedly, the number of unique homozygous substitutions observed in our control plants was similar to the number in the FN lines or transgenic plants. This implies most of the unique homozygous substitutions were likely due to spontaneous mutation rather than an increased mutation rate in the generation of mutagenesis or transformation. In terms of single nucleotide substitutions, our result implies minimal difference in the safety risks in any of the three germplasm classes. This result is in contrast to studies of tissue culture in rice that suggest a significantly higher number of induced homozygous substitutions and associated mutation rate (Miyao *et al.* 2012; Zhang *et al.* 2014). A number of confounding factors might affect these incongruities including differences in the species examined, SNP calling methods and thresholds, adjustments for intra-cultivar heterogeneity, FN dosage or tissue culture conditions and timeline, the inclusion of a control plant, and the number of lines sampled.

Based on our data, it appears the use of FN mutagenesis can produce profound new SV events and may slightly increase the number of single nucleotide substitutions.

Tissue culture/transformation methodologies can also produce new SV and possibly increase the nucleotide substitution rate. Furthermore, the number of SV and single nucleotide polymorphisms existing as standing variation in soybean cultivars dwarfs the induced variation observed in both FN and transformed plants. These findings are noteworthy but it is unclear how broadly they can be applied. All of the transgenic plants in this study were obtained from *Agrobacterium*-mediated transformation; further work would test other transformation techniques such as biolistic-based methods. Similarly, FN irradiation was the only mutagenesis system tested; other mutagens (EMS, ENU, etc.) would likely induce different mutational profiles. Furthermore, a deeper sampling of mutated and transformed plants, perhaps among different plant species, would be required to generalize the SV and nucleotide trends observed. Detailed sequence analysis of specific transgene loci did identify a small number of intermediate-sized deletions adjacent to transgenes, but there was no systematic attempt to detect intermediate-sized (1-2,000 bp) deletions/duplications genome-wide. Additional variants have also been reported to exist in FN (Bolon *et al.* 2014) and transgenic lines (Tax and Vernon 2001; Cheng *et al.* 2008; Clark and Krysan 2010; Majhi *et al.* 2014) but were not assessed within this dataset, including inversions and translocations, as well as epigenetic or transcriptional perturbations. Lastly, soybean is a palaeopolyploid species. It is likely that true polyploid (or true diploid) species may exhibit differential tolerance or lack of tolerance to the type of genetic perturbations associated with these technologies.

Conclusions

The total findings of this study help to inform the discussion currently surrounding the unintended consequences of genetic transformation in crop improvement (Weber *et al.* 2012; Schnell *et al.* 2015). First, the frequency of induced SV events appears to be low, particularly in comparison to the frequency of those induced by FNs. Additionally, these rare SV events are likely indistinguishable from other spontaneously occurring SV or those already present in the existing germplasm. As demonstrated by the genetic variability in the no-phenotype FN lines, SV are not always associated with novel or noticeable phenotypic traits. Therefore, the speculated risk of unintended genetic consequences in tissue culture/transformation merit only as much consideration as given to variation arising spontaneously, through traditional breeding practices, or other genetic variation induction methods.

MATERIALS AND METHODS

Plant Materials and Genetic transformation

The plant materials comprising the inter-cultivar and FN classes included in this study have been previously described (Anderson *et al.* 2014; Bolon *et al.* 2014). Briefly, the inter-cultivar group consists of 41 soybean accessions used as parents in developing the SoyNAM population. The FN population was developed in the background of the variety 'M92-220'(Bolon *et al.* 2011) derived from the 2006 Crop Improvement Association seed stock of variety 'MN1302'(Orf and Denny 2004). To protect against a sampling bias that favors high rates of structural variation, only FN treated plants with no known mutant phenotypes were included in this study. This group, known as the "no phenotype" sub-sample, includes 35 lines descended from 35 unique M_1 individuals that were treated with either 4, 16, or 32 Gy of FN radiation (Bolon *et al.* 2014).

Genetic transformation using *Agrobacterium rhizogenes* followed published methods (Paz *et al.* 2006; Curtin *et al.* 2011). Each plant was confirmed to be transgenic based on PCR analysis and survival on selective (herbicide-treated) medium. The five T_1 soybean individuals were from unique transformation events. The constructs for these transformations included a zinc finger nuclease, transcription activator-like effector nuclease, GFP and RNAi hairpin, mPing-Pong transposon, and a magnesium chelatase RNAi hairpin. These transformations were in a 'Bert' cultivar (Orf and Kennedy 1992) background (subline 'Bert_MN01') or a 'Williams 82' subline ('Wm82_ISU_01')(Bernard and Cremeens 1988; Haun *et al.* 2011). The 'Bert_MN01' subline (referred to as 'Bert' throughout this study) was derived from a single Bert individual to reduce heterogeneity between transformed lines. The 'Wm82_ISU_01' subline (referred to as 'Williams 82' throughout this study) was derived from a single Williams 82 individual and is the nearest known match to the soybean reference genome assembly version 1.0 (Schmutz *et al.* 2010; Haun *et al.* 2011).

Comparative Genome Hybridization

The CGH data for all comparisons used in this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). The data for the in inter-cultivar, FN, and transgenic plant comparisons can be found as accession numbers GSE56351, GSE58172, and GSE73596, respectively.

As with previous CGH analyses (Anderson *et al.* 2014; Bolon *et al.* 2014), the DEVA software algorithm SegMt was used to generate raw data and identify segments in the transgenic plants. Transgenic lines were labeled with Cy3 and the appropriate reference individual (Bert or Williams 82) was labeled with Cy5. Program parameters were: minimum segment difference = 0.1, minimum segment length (number of probes) = 2, acceptance percentile = 0.99, number of permutations = 10. Spatial correction and qspline normalization were applied. The resulting segments were processed based on their $\log_2$ ratio mean. Segments that exceeded the upper threshold were considered "UpCNV." Segments that were less than the lower threshold were considered "DownCNV." The upper threshold of 0.3484 and lower threshold of -0.5257 were based on empirical data from hemizygous deletions and duplications in eight previously characterized FN lines (Table S6) (Bolon *et al.* 2014). A custom Perl script calculated the number genes overlapping these significant segments. Minimum segment length was adjusted to three probes to account for noise seen in control arrays. Structural variants in the transgenic lines were further investigated through visual inspection, to identify any obvious SVs that were not detected by the threshold based pipeline.

Next, SV attributable to intra-cultivar heterogeneity were removed, as has been done in the previous studies (Anderson *et al.* 2014; Bolon *et al.* 2014). Intra-cultivar heterogeneity was seen as significant segments of the exact same location occurring in multiple lines. By overlaying the raw CGH data of the four transgenic lines in the Bert background, heterogeneous SV in the Bert cultivar were removed. A similar method was used to filter out heterogeneity in the transformed Williams 82 background. The comparison array in this case was Williams (the backcross parent in Williams 82 (Bernard and Cremeens 1988)) also hybridized to Williams 82. Any identical SV event discovered in both Williams and transformed Williams 82 was considered heterogeneity and removed.

The CGH platform, methods, and filtering steps of the inter-cultivar and FN data have been previously described (Anderson *et al.* 2014; Bolon *et al.* 2014). Notably, the SV detected in the inter-cultivar variation study were all cross validated with resequencing data and conservative thresholds. For all CGH arrays, test genotypes were labeled with Cy3 and the appropriate reference individual was labeled with Cy5 in all hybridizations (Table S2).

Visual displays of the CGH data were created using Spotfire DecisionSite software. Table S7 provides a list of soybean lines chosen for analysis, corresponding publication, and hybridization reference. Our previous study (Anderson *et al.* 2014) of inter-cultivar variation concluded SV affected 1528 genes by assessing CNV on a gene-by-gene cross-validated basis across all 41 SoyNAM genotypes. We conservatively converted this to SV genes per genotype using the CGH thresholds from the study and probe-based $\log_2$ ratio score for each of the 1528 genes. FN data came from the “no phenotype” class of 35 lines, as described above (Bolon *et al.* 2014). Only SV overlapping genes were included in segment size summaries in all three genotypic classes.

Confirming Novel SV

PCR was used to confirm structural variants found via CGH in the transgenic lines. PCR and Sanger sequencing across breakpoints was able to confirm the four CGH observed events. Confirmed events and internal primers were used for genotyping these structural variants in additional lines. Primer sequences are provided in Table S8. In three of these lines siblings and offspring of the transgenic plants were genotyped to confirm the SV were heritable. The events were confirmed not to be intra-cultivar heterogeneity by PCR-genotyping 47 untransformed lines (either in the corresponding 'Bert' or 'Williams 82' background) at these three loci. Furthermore, the SoyNAM parents as well as cultivars 'Archer', 'Minsoy', and 'Noir1' were also PCR-genotyped with the breakpoint and internal primers to test for novelty of the SV events.

Analyzing Transgene insertion sites

Transgene integrations were analyzed using TAIL-PCR, Southern blot, and resequencing data. Southern blots used a BAR gene probe to detect the number of T-DNA insertions in the lines tested. TAIL-PCR(Singer and Burke 2003) was used to detect T-DNA locations in WPT_384-1-1, WPT_389-2-2 and WPT_301-3-13. Transgene insertion sites and counts were also determined by resequencing according to steps one through six outlined by (Srivastava *et al.* 2014). Briefly, raw paired-end reads were aligned using Bowtie2 to the transgene sequence between the left and right border and the orphaned mapped reads were then aligned to the host soybean genome. The resulting putative transgene integration locations were filtered on prior knowledge of homology between components of the transgene (i.e. Gmubi promoter, RNAi hairpin targets, and their paralogs) and the genome. The location of the mapped orphaned reads, read depth coverage, and paired-end read spacing were further used to detect SV induced locally to transgene insertions. Integrated Genome Viewer (IGV) version 2.3.52 was used to visualize alignment results (Thorvaldsdóttir *et al.* 2013).

Sequence Handling, Alignment, and Calling of Nucleotide Substitutions

The sequence read data from the ten fast neutron plants analyzed in this study, along with the parent line of the population (cv. 'M92-220'), are deposited in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP036841. The sequence read data from the two transgenic plants, along with two individuals of the parent line (cv. 'Bert'), and the cultivars 'Archer' and 'Noir 1' are deposited in the Sequence Read Archive under accession number SRP063738.

To determine the relative rates of base substitution due to FN mutagenesis, we used resequencing data from a subset of the FN population reported in (Bolon *et al.* 2014). These lines had associated phenotypes but the number of genes affected by SV was similar to those in the no-phenotype class (see Table S4) suggesting they were an acceptable comparison. We additionally sequenced two transgenic plants and two controls to estimate the base substitution rate and localize T-DNA insertion sites. See Figure S11 for the transgenic resequencing data analysis pipeline. All lines were sequenced with Illumina 100 bp paired end reads.

FastQC version 0.11.2 was used on initial read data and after any modifications to sequence data to ensure that tools were used properly and the data was of acceptable quality for downstream applications (Andrews 2010). Forward and reverse reads were treated separately, and then resynchronized for alignment using resync.pl (Riss util version 1.0, http://msi-riss.readthedocs.org/en/latest/software/riss_util.html). Cutadapt version 1.6 was used to remove adapter sequences using –b to specify both adapter sequences (GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGT-ATGCCGTCTTCTGCTTG, AATGATACGGCGACCACCGAGATCTACACTCTTTCCC-TACACGACGCTCTTCC-GATCT) where NNNNNN specifies the unique 6bp sequence attached to samples when multiplexing. Sequence artifacts (low-complexity reads) were removed using fastx artifacts filter (Fastx toolkit version 0.0.14). Read quality was further filtered using fastq quality trimmer in the fastxtoolkit. Bases with phred quality of less than 20 were removed, and reads that were shorter than 30 bp after trimming were discarded.

We chose to align reads to the reference with two different read mapping programs, BWA mem (v. 0.7.10)(Li and Durbin 2009a), and Bowtie2 (v. 2.2.4)(Langmead and Salzberg 2012). BWA mem alignments allowed for more accurate single base substitution calls, and Bowtie2 produces alignments more suitable for confirming CGH-identified SV. For BWA mem, mismatch penalty was set to 6 (-B 6), which allows for approximately seven high-quality mismatches per read. Bowtie2 alignments were produced with default parameters. In both cases, reads were mapped to the *Glycine max* assembly version 1 (Schmutz *et al.* 2010). Read cleaning and post-alignment filtering resulted in a realized mean coverage of 35x for the FN mutagenized lines, and 20x for WPT_389-2-2, and 21x for WPT_391-1-6.

Genotype calls for all sites were generated with the UnifiedGenotyper in the Genome Analysis Tool Kit (GATK) version 3.3 (DePristo *et al.* 2011). Pairwise comparisons of soybean varieties typically identify over one-million single base substitutions (Lam *et al.* 2010; Zhou *et al.* 2015). This BWA mem resequencing and SNP detection pathway identified 1,110,325 substitutions between genotype 'Archer' and the 'Williams 82' reference genome sequence, and 1,904,061 substitutions between genotype 'Noir 1' and 'Williams 82'. These findings served as a control to demonstrate our analysis pipeline identify similar polymorphism counts as have been previously reported in soybean studies.

We then applied a set of filtering criteria to look at only unique substitutions across the most confidently called portions of the genome. This excluded sites with less than five reads per sample, sites that were monomorphic for the reference base, sites with heterozygous or missing calls, and sites with a homozygous alternate base call in more than one individual. Applied together, these filtering criteria produce variant calls that are homozygous private differences from reference. It is important to note our filtering criteria assumed mutations at a single base position will only be observed once. A large section in FN line 07 on Chromosome 12 between 10 and 23 Mb was found to contain a disproportionate number of substitutions. CGH results from other FN lines (Bolon *et al.* 2014), not included in this sample, suggest this region is heterogeneous in the 'M92-220' cultivar. We therefore excluded this region of 183 substitutions when analyzing FN line 07. The observed transition:transversion ratios were too variable between lines to compare to previously reported ratios in FN mutagenesis (Belfield *et al.* 2012).

Circos plots (Krzywinski *et al.* 2009) were generated using 2d tile data tracks plotting unique substitutions detected, previously published FN-induced SV (Bolon *et al.* 2014), detected transformation-induced SV, and T-DNA mapping results. Scripts to perform data handling and analysis are available at https://github.com/TomJKono/Unintended_Consequences.

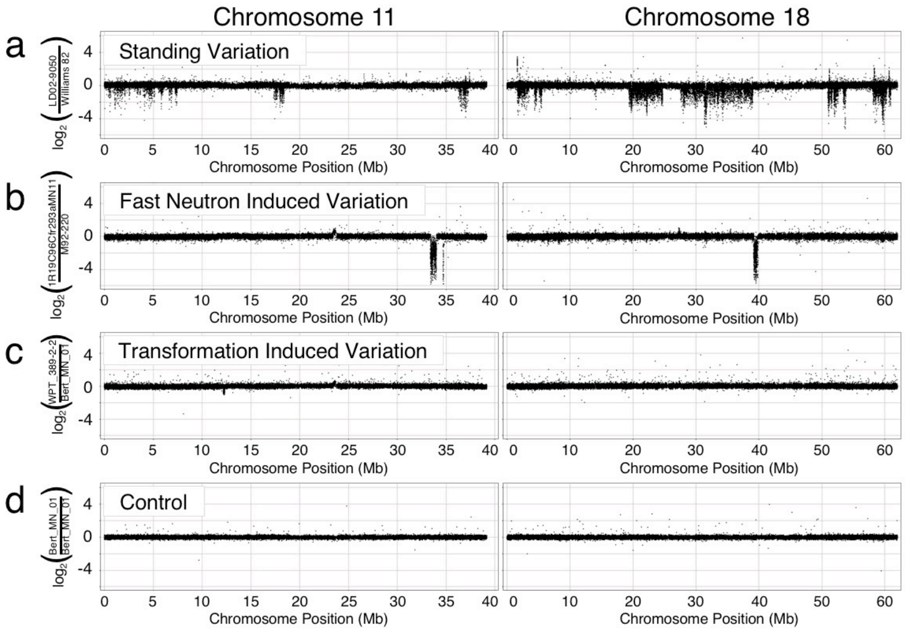


Figure 1. Visual comparison of CGH data for individuals from the three germplasm classes and control. Each black dot represents a single probe and its $\log_2$ ratio score. All genotypes are only showing data from chromosome 11 on the left and chromosome 18 on the right. (a) The standing variation detected as inter-cultivar by CGH on line LD02-9050 shows high noise but distinct SV. This line has 254 putatively deleted or duplicated genes across the genome when comparing to ‘Williams 82’. (b) The CGH on fast neutron line 1R19C96Cfr293aMN11 shows low noise throughout and one SV segment on both chromosomes. This line has 124 putatively deleted or duplicated genes across the genome when compared to ‘M92-220’. (c) The CGH on transgenic plant WPT_389-2-2 shows relatively little noise and one true SV on chromosome 11 compared to ‘Bert’. This line has 4 genes deleted across the genome when comparing to ‘Bert’. (d) The control CGH on ‘Bert_MN_01’ shows only small amounts of background technical noise and did not detect any deleted or duplicated genes across the genome.

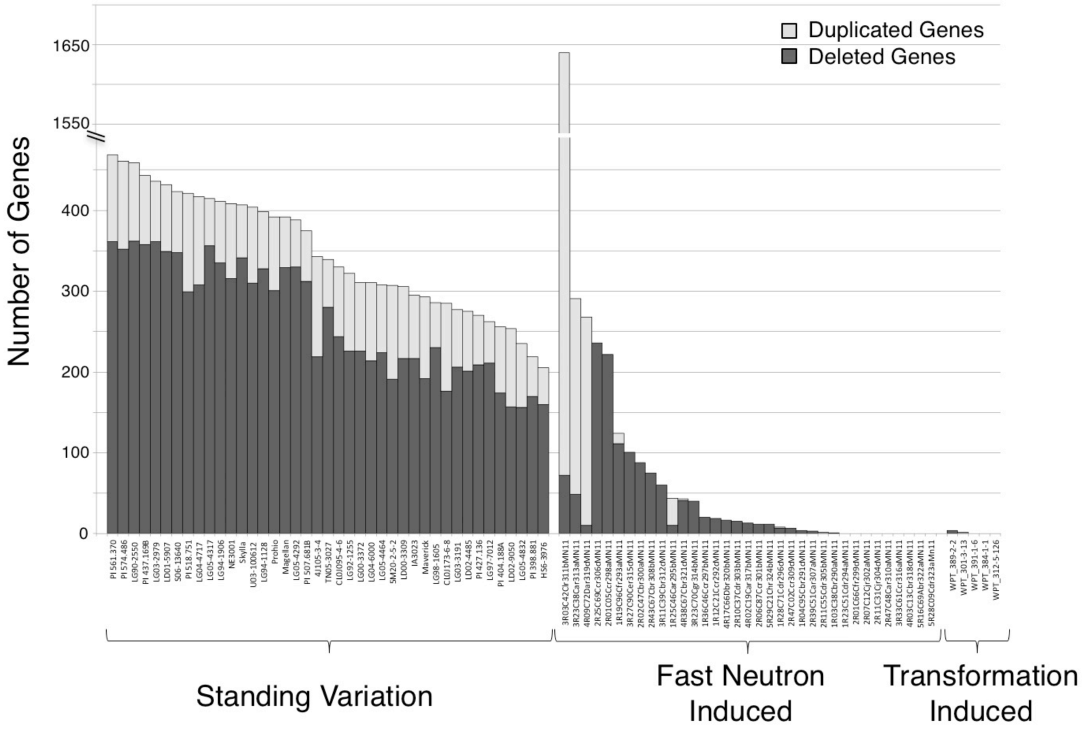


Figure 2. Distribution of genic SV as standing variation in diverse cultivars (41 SoyNAM parents), induced by fast neutron mutagenesis (35 FN lines with no obvious mutant phenotypes), or induced by the transformation process (five lines with unique constructs). Each column is a single genotype. Light gray bars represent "Duplicated Genes," those overlapping putatively duplicated regions, and dark gray bars represent "Deleted Genes," those overlapping putatively deleted regions.

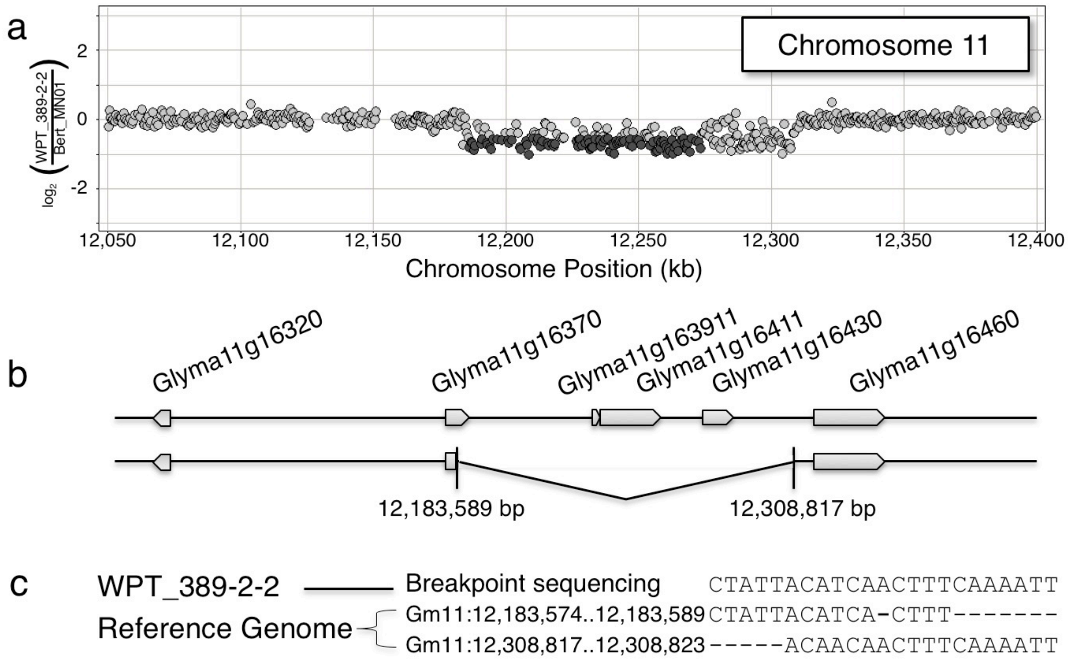


Figure 3. A novel deletion on chromosome 11 in transgenic line WPT_389-2-2. (a) Plot of CGH data for the transgenic line versus 'Bert', zoomed in on the chromosome 11 deletion seen in Figure 1*C*. Probes are plotted as dots corresponding to the $\log_2$ ratio from the CGH array. Dark gray dots represent probes within significant segments that exceed the empirical threshold. Even with the extremely low detection threshold, part of this deletion could not be verified via CGH alone necessitating visual inspection and breakpoint sequencing. (b) Graphical interpretation of the hemizigous deletion found in WPT_389-2-2. (c) Sequence data from the breakpoint junction shows moderate homology on either end of the breakpoint.

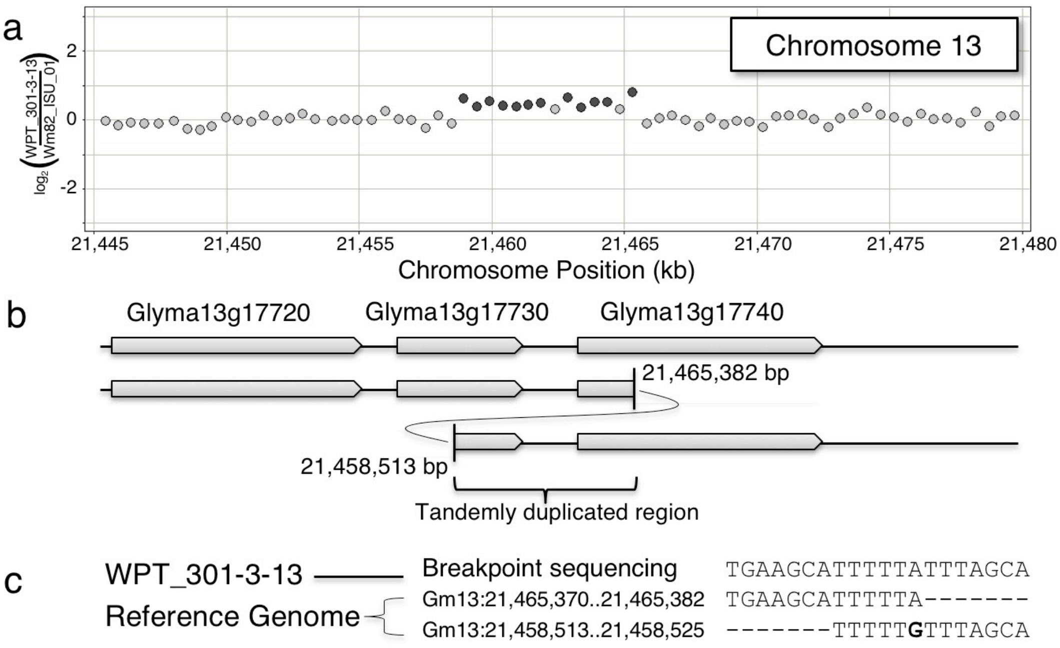


Figure 4. A novel duplication on chromosome 13 in transgenic line WPT_301-3-13. (a) Plot of CGH data for the transgenic line versus 'Williams 82', zoomed in on the chromosome 13 duplication. Probes are plotted as dots corresponding to the $\log_2$ ratio from the CGH array. Dark gray dots represent probes within significant segments that exceed the empirical threshold. (b) Graphical interpretation of the heterozygous duplication found in WPT_301-3-13. (c) Sequence data from breakpoint junction shows five base pairs of homology on either end of the breakpoint. This duplication included a portion of Glyma13g17730 and a portion of Glyma13g17740, but did not include any complete genes.

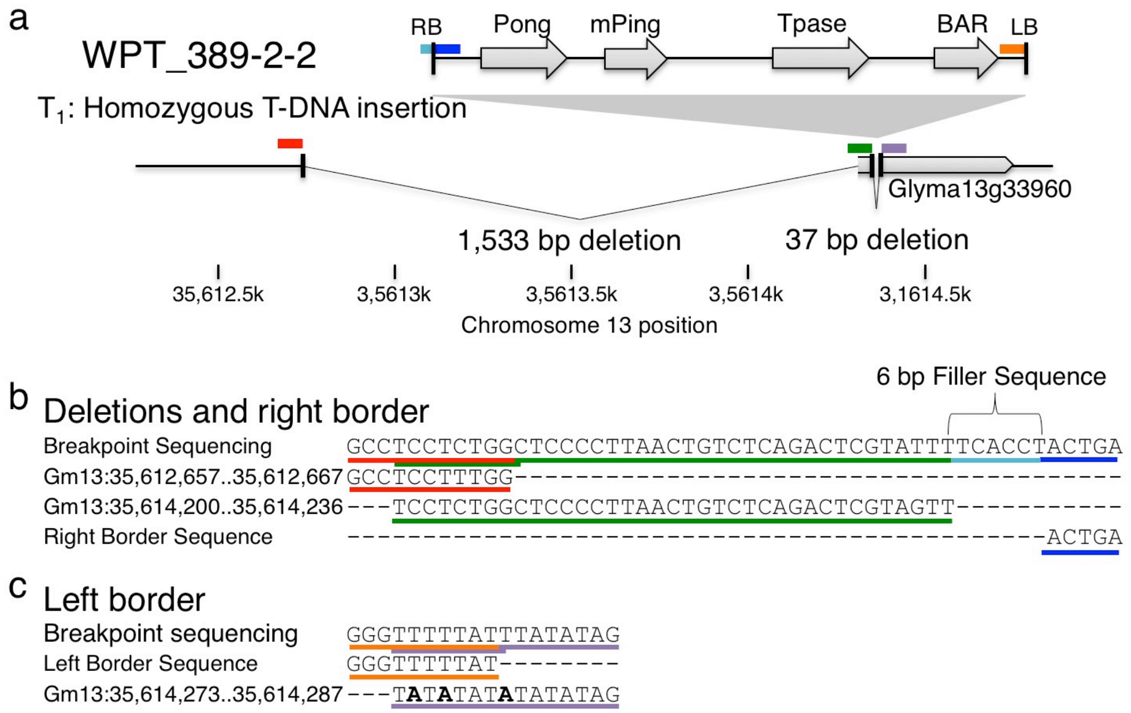

a
WPT_389-2-2
T1: Homozygous T-DNA insertion
RB
Pong
mPing
Tpase
BAR
LB
Glyma13g33960
1,533 bp deletion
37 bp deletion
35,612.5k
3,5613k
3,5613.5k
3,5614k
3,1614.5k
Chromosome 13 position
b
Deletions and right border
6 bp Filler Sequence
Breakpoint Sequencing GCCTCCTCTGGCTCCCCTTAACTGTCTCAGACTCGTATTTTCACCTACTGA
Gm13:35,612,657..35,612,667 GCCTCCTTTGG--
Gm13:35,614,200..35,614,236 ---TCCTCTGGCTCCCCTTAACTGTCTCAGACTCGTAGTT-----------
Right Border Sequence --ACTGA
c
Left border
Breakpoint sequencing GGGTTTTTATTTATATAG
Left Border Sequence GGGTTTTTAT--------
Gm13:35,614,273..35,614,287 ---TATATATATATATAG

Figure 5. Transgene insertion locus and induced homozygous deletions in genome of WPT_389-2-2. **(**a) Graphical interpretation of the transgene orientation and induced deletions at this locus. The transgene insertion contains four primary elements between the left and right borders: Pong, mPing, Tpase, and BAR on chromosome 13 in line WPT_389-2-2. Colored lines correspond to the breakpoint sequence results. (b) Results of breakpoint sequence data spans from the genome (red), across the 1,533 bp deletion back into genome space (green), across filler sequence (light blue) and into the T-DNA right border (dark blue) and (c) from the T-DNA left border (orange) into the genome (purple). Microhomology occurs across the large deletion and between the left border and the genome. This T-DNA insertion appears to have induced a local 1,533 bp deletion, a small insertion of filler sequence, and an additional 37 bp deletion in the process of integration.

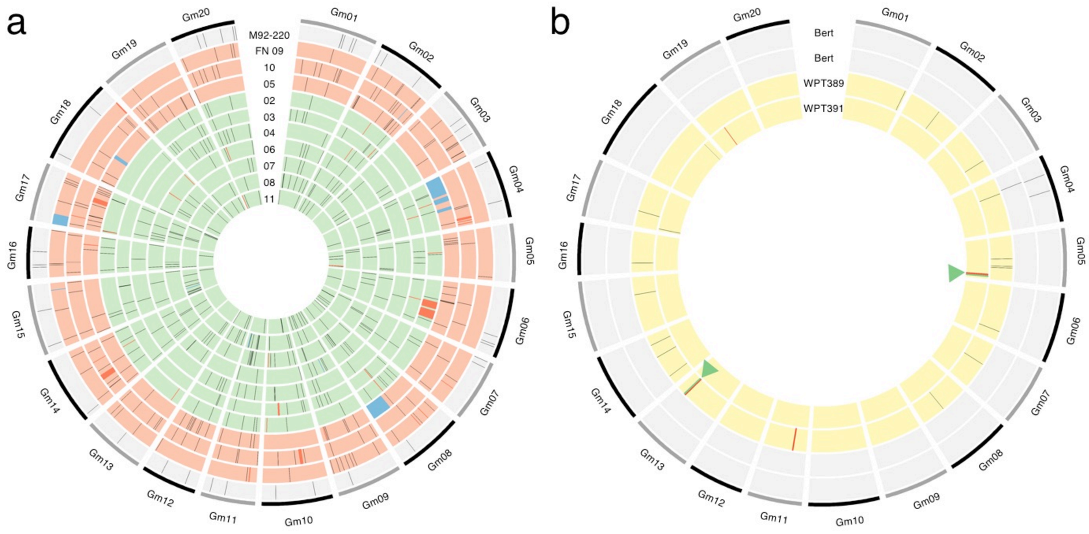


Figure 6. Genome wide view of induced variation detected through CGH and resequencing. Black bars are substitutions, blue bars are duplications, and red bars are deletions. Regions were filtered for heterogeneity; therefore only plant-specific variation is shown. (a) Fast neutron lines, including the parent ‘M92-220’ (outer ring) and FN02-FN11 (inner rings). Background is shaded according to fast neutron irradiation dosage: gray is non-irradiated parent ‘M92-220’, red is 32 Gy (FN 09, 05 and 10), and green is 16 Gy (FN 02, 03, 04, 06, 07, 08, and 11). Variation detected in ‘M92-220’ is likely due to spontaneous mutation rather than a byproduct of heterogeneity. (b) Unique genetic variation in two different sequenced ‘Bert’ parent individuals (gray background), and transgenic plants WPT_391-1-6 and WPT_389-2-2 (yellow backgrounds). Transgene insertion sites are noted by green arrows and bars. Variation detected in ‘Bert’ is likely due to natural spontaneous mutation. Overall, fast neutrons appear to induce more SV and substitutions than transformation in these plants.

Chapter 4: Environmental association analyses identify candidates for abiotic stress tolerance in *Glycine soja*, the wild progenitor of cultivated soybeans[4]

Natural populations across a species range demonstrate population structure owing to neutral processes such as localized origins of mutations and migration limitations. Selection also acts on a subset of loci, contributing to local adaptation. An understanding of the genetic basis of adaptation to local environmental conditions is a fundamental goal in basic biological research. When applied to crop wild relatives, this same research provides the opportunity to identify adaptive genetic variation that may be used to breed for crops better adapted to novel or changing environments. The present study explores an *ex situ* conservation collection, the USDA germplasm collection, genotyped at 32,416 SNPs to identify population structure and test for associations with bioclimatic and biophysical conditions variables in *Glycine soja,* the wild progenitor of *Glycine max* (soybean). Candidate loci were detected that putatively contribute to adaptation to abiotic stresses. The identification of potentially adaptive variants in *ex situ* collection may permit a more targeted use of germplasm collections.

[4] This chapter is the result of collaborative research. Co-authors include: Thomas J. Y. Kono, Robert M. Stupar, Michael B. Kantar, and Peter L. Morrell. The author of this dissertation contributed to designing and performing the experiments, environmental association analysis, parsing results, exploring homology, drafting the manuscript, developing figures and tables, and editing. TJYK and MBK contributed notably in the areas of association analysis design, SPA, and population genetic methods. RMS and PLM helped conceive and design the study and supervised the analysis. All authors reviewed, commented, and approved the manuscript. This work was submitted to G3: Genes, Genomes, and Genetics in August 2015 and was under review at of Nov 2015. Supplemental data can be found in Appendix 2.

INTRODUCTION

It has long been observed that individuals of the same species from different local environments have distinct phenotypes. Individuals tend to have higher fitness in their environment of origin, being adapted to this locality (Fournier-Level *et al.* 2011). Local adaptation is particularly important in plants, as sessile organisms cannot relocate to more hospitable environmental conditions (Tiffin and Ross-Ibarra 2014). Environmental association is particularly appealing in studies of crop wild relatives, as the variation identified may then be tested for targeted crop improvement. Because cultivars are typically derived from only a limited subset of wild progenitors (Harlan *et al.* 1973), environmental association studies of crop wild relatives have the potential to uncover adaptive variants that do not occur in current cultivars.

A number of approaches have been developed to identify genetic variants contributing to local adaptation. Lewontin and Krakauer (Lewontin and Krakauer 1973) first proposed the comparisons of subpopulations to identify loci with large allele frequency differences as measured by F_{ST} (Wright 1949), an approach that became known as the “Lewontin and Krakauer Test.” A number of criticisms have been leveled against the Lewontin and Krakauer Test including high variance in F_{ST} (Nei and Maruyama 1975; Robertson 1975) and sensitivity to differences in sample sizes (Weir and Cockerham 1984; Hudson *et al.* 1992). Despite these limitations, simulation studies suggest the Lewontin and Krakauer framework provides a useful means of identifying potentially adaptive variants (Beaumont and Balding 2004; Beaumont 2005). Newer approaches operate in slightly different frameworks to identify allele frequency gradients or even association with environmental variation. For example, Spatial Ancestry Analysis (SPA)(Yang *et al.* 2012) is appropriate for sampling of individuals across continuous geographic space and environmental gradients. A continuous function of allele frequency is estimated and projected onto geographic space, and loci showing steep gradients in allele frequency are interpreted to be associated with locally adaptive variation (Yang *et al.* 2012).

Mixed model association mapping, often used in studies of phenotypic variation (Lipka *et al.* 2015), can also be used in environmental association studies (Yoder *et al.* 2014). In this framework, environmental data are treated as "phenotypes" and the genetic data are queried for variants most strongly associated with these environmental phenotypes (Eckert *et al.* 2010; Yoder *et al.* 2014). Public repositories of global bioclimatic (WorldClim) (Hijmans *et al.* 2005) and biophysical (soils) variables (ISRIC) (Hengl *et al.* 2014) with up to 1 km resolution are currently available for association studies.

Exploring local adaptation in crop wild relatives has important agricultural implications. Crop wild relatives are often a source of novel genetic variation for plant breeding (McCouch *et al.* 2013; Khoury *et al.* 2015). Much of the introgression of adaptive variation from wild relatives has, to date, involved crosses with single accessions containing favorable characteristics (e.g., resistance to a particular pathogen). Exploration on a population scale with the inclusion of environmental data has the potential to reveal variation linked to adaptations to abiotic stress tolerance. One concern is the detected abiotic stress alleles available may be limited to the wild niche the crop wild relative resides in. Crop wild relatives often inhabit different ecological niches from the domestic material (Khoury *et al.* 2015). These niches can be broader or narrower depending on the environmental variable being examined.

Glycine soja, the wild progenitor of cultivated soybean, is native to East Asia with a broad distribution in China, Japan, Korea, and Russia (Li *et al.* 2010). There is extensive environmental variation across its native range, with altitude ranging from sea level to ~1400 m, yearly precipitation ranging from 300-3400 mm, and mean annual temperature ranging from -3.1 to 18.2º C. This environmental range is quite similar to that found in major soybean cultivation regions of North America. In North America, soybean is cultivated in areas with altitude ranging from sea level to ~900 m, yearly precipitation ranging from 400-1800 mm, and mean annual temperature ranging from 1.3 to 20.5º C. Given this environmental similarity, and the ability to cross *Glycine soja* and cultivated soybean, detected associations can be readily tested and implemented in soybean breeding programs.

In this study, we examined population structure, environmental associations, and allele frequency gradients in 533 accessions of *Glycine soja*. The sampled accessions are derived from the USDA GRIN soybean germplasm collection and were genotyped with the SoySNP50K genotyping platform (Song *et al.* 2015). Environmental association, SPA, and F_{ST} outliers were explored, identifying loci that may be useful in targeted improvement of abiotic stress tolerance in soybean.

MATERIALS AND METHODS

Genetic Data Acquisition

Genotype data from the SoySNP50K platform (Song *et al.* 2015) were downloaded from SoyBase (Grant *et al.* 2010) for all available *G. soja* accessions. Among those with latitude and longitude coordinates, accessions were removed if they had greater than 10% missing data, were genetically redundant, or were geographic outliers from Taiwan and Northern Russia, yielding 533 accessions from 273 unique sampling locations. Ambiguous and heterozygous SNP calls were treated as missing data due to the low outcrossing rate (~3%) in *G. soja* (Kuroda *et al.* 2006; Guo *et al.* 2012). Monomorphic sites were also removed leaving 32,416 polymorphic SNPs. These SNPs were distributed throughout the euchromatic and pericentromeric regions and spaced at an average of ~8.6 kb and ~45 kb, respectively. A list of the accession (Plant Introduction or PI) numbers and geographic origins of the *G. soja* accessions used in this study is available in Table S1, and a map of our sampled accessions is shown in Figure 1. The physical positions of the SoySNP50K SNPs (Song *et al*. 2013) were mapped into the second genome assembly 'Glyma.Wm82.a2' (http://www.soybase.org/), where the SNP query sequences were aligned with Bowtie 2 (Langmead and Salzberg 2012). The resulting SAM (Li *et al.* 2009) file was parsed with a custom Python script to extract the SNP position on the version 2 assembly.

Bioclimatic and Biophysical variables

Latitude and longitude coordinates associated with *G. soja* sampling locations were used to query the WorldClim database for 68 variables, including bioclimatic variables based on yearly, quarterly, monthly temperature and precipitation data as well as altitude data at a resolution of 30 arc-seconds (approximately 1km grids)(Hijmans *et al.* 2005). The sampling locations (longitude and latitude) were also used to query the ISRIC database (World Soil Information database, http://soilgrids1km.isric.org) for seven biophysical variables (pH x 10 in H_2O, percent sand, percent silt, percent clay, bulk density in kg/cubic-meter, cation exchange capacity in cmolc/kg, and organic carbon content (fine earth fraction) in permilles) at a resolution of 30 arc-seconds. Soils data was also grouped into two classes: topsoil (from 0-30 cm) and subsoil (from 30-200 cm), resulting in fourteen soil variables. Classes were created by averaging the appropriate depths from the six depths available in the ISRIC database: 2.5 cm, 10 cm, 22.5 cm, 45 cm, 80 cm, and 150 cm (Hengl *et al.* 2014). Both of these represent the highest resolution available for these data. Principle component analysis (PCA) on the bioclimatic and biophysical variables (first scaled to a mean of 0 and standard deviation of 1) was conducted using the prcomp function in R (R Development Core Team 2011). Pearson correlations between bioclimatic and biophysical variables were also calculated in R. Boxplots for each scaled bioclimatic and biophysical variable were created based on *G. soja* localities to confirm variability in these data (Figure S1).

Population Structure, Allelic Composition, and Linkage Disequilibrium Measures

Genetic assignment analysis was used to identify population structure in the sample using a Bayesian Monte Carlo Markov Chain (MCMC) algorithm implemented in STRUCTURE (Pritchard *et al.* 2000). The number of clusters (*K*) from 2 to 5 was explored using a model with uncorrelated allele frequencies and no admixture between clusters, parameters that reflect a high degree of observed allele frequency differentiation among populations and no prior evidence of admixture in *G. soja*. Runs for each *K* value were replicated 10 times, with 10,000 burn-in steps and recorded for 10,000 subsequent steps. STRUCTURE assignments were visualized with the CLUMPPAK server (Kopelman *et al.* 2015). PCA was also used to explore population structure using the SNPRelate package (Figure S2) (Zheng *et al.* 2012).

Allele frequency differentiation (F_{ST}) was estimated among populations identified through genetic assignment (STRUCTURE). Theta (Θ), the variance-based F_{ST} estimate of Weir and Cockerham (Weir and Cockerham 1984), was estimated in the R 'hierfstat' package. The private allele richness of populations was calculated with the rarefaction approach ADZE to account for differences in sample size (Szpiech *et al.* 2008). For visualization, F_{ST} was averaged in sliding windows, with a window size of 5 and a step of 3 SNPs. A Mantel test was conducted to explore isolation by distance utilizing great circle distance between geographic locations and pairwise genetic distance using the 'vegan' package in R.

SPA was used to detect loci showing steep gradients in allele frequency (Yang *et al.* 2012). SNPs were designated outliers if they fell above the 99.9th percentile of the distribution of SPA selection scores. SPA should better deal with isolation by distance as it incorporates geographic and genetic gradients in search of local clines, unlike a search of F_{ST} outliers which are predicated on user-defined population structure (Yang *et al.* 2012).

The extent of linkage disequilibrium (LD) in the sample was calculated with the 'LDheatmap' package in R. LD as D′ (Lewontin 1964) was calculated between all pairwise combinations of markers on each chromosome. LD decay over physical distance was estimated using the exponential regression method (Abecasis *et al.* 2001). Due to the strong difference in recombination rate between pericentromeric regions and euchromatic regions (Lee *et al.* 2015), we treated these regions separately. For calculating the decay curves, we used 2.39 cM/Mb and 3.59 cM/Mb for pericentromeric and euchromatic regions, respectively, based on median adjacent-SNP recombination rate from the genetic map of Lee *et al.* (2015).

Environmental Association Mapping

Mixed-model association as implemented in Tassel (5.0v) (Bradbury *et al.* 2007; Zhang *et al.* 2010) was used to test for associations between individual SNPs and bioclimatic and biophysical variables. To identify the appropriate association model, the following models were explored: the naïve model with no control of population structure, a model using the Q-matrix from STRUCTURE, a model using a kinship matrix (*K*-matrix), a model using both a *K* matrix and a *Q*-matrix, and models also integrating latitude or latitude and longitude as covariates. Quantile-Quantile (qq) plots were examined for each model and the genomic inflation parameter lambda (Λ) was calculated (Figure S3). The final model utilized the bioclimatic/biophysical variable as the response and genotype as a fixed effect, *K*-matrix as a random effect, and latitude as a covariate. This was the simplest model (least covariates) with a Λ near 1 (Figure S3). The use of latitude or latitude and longitude as covariates resulted in nearly the same Λ and therefore only latitude was used as a covariate to prevent any potential over correcting. Utilizing latitude as a covariate also likely addressed possible confounding by flowering time. The sample was not divided into clusters for separate environmental associations because the Mantel test suggests isolation by distance as the primary driver of population structure. Additionally, the sample locations are distributed across the entire range, dividing into the geographic clusters would reduce power to detect associations by decreasing the number of environments sampled and the number of individuals tested.

A cutoff of the 0.01% most extreme *p*-values were explored as candidates for each environmental association resulting in three significant markers. This strict threshold was chosen to focus the analysis on a minimum number of large effect QTL and limit the number of false positives. While such a strict threshold likely excludes many true positive associations of small effect, these initial significant associations should be more impactful if tested and implemented in soybean breeding programs. The qqman R package (Turner 2014) was used to plot the association results.

Candidate Characterization

SNPs identified as outliers through the environmental association mapping, SPA, or F_{ST} approaches were examined for functional annotation using SoyBase (www.soybase.org) (Grant *et al*, 2010). This database provided access to minor allele frequency (MAF) within landrace, elite lines, and *G. soja* panels, based on data from the SoySNP50K development study (Song *et al.* 2013). Further molecular information, including genic context, nearby annotated genes, and a gene's Arabidopsis ortholog (TAIR10 best hit according to Soybase), was also assessed. Outliers were explored for enrichment in euchromatin, 3′ UTR, 5′ UTR, coding sequence (CDS), and intronic regions. Significance of enrichment was assessed by creating a 99% confidence interval around the proportion of SNPs that were found in each category as calculated by bootstrap sampling the number of SNPs in each category 1000 times. The scripts and small input files used to filter SNPs, run STRUCTURE, calculate F_{ST}, calculate LD, and generate figures are publicly available in the GitHub repository located at https://github.com/MorrellLAB/Soja_Env_Association.

RESULTS

Population Structure

From the GRIN soybean germplasm collection, 533 accessions of *Glycine soja* were used in this study. This subset had longitude and latitude data, were not genetically identical to another accession, and had less than 10% missing data. These accessions represent 273 unique sampling localities across East Asia (Figure 1). Genetic data included a filtered list of 32,416 polymorphic SNP markers from the GRIN soybean germplasm SoySNP50K genotyping efforts (Song *et al*. 2015). Genetic assignment was assessed at *K*=2 to *K*=5. Genetic assignment at *K* = 2 divided accessions primarily east and west of the Sea of Japan. At *K* = 3, the Japanese Archipelago samples form a distinct cluster, and the mainland samples were split into a northern and southern cluster. With *K* = 4, the samples located on the Korean Peninsula began to separate from mainland Asia, forming a unique cluster, or more infrequently the Japanese samples subdivided into two clusters. For *K* = 5, the Japan cluster separated into two distinct northern and southern subpopulations. The majority of analyses reported here are based on *K* = 3, which was identified as the optimum number of clusters (Evanno *et al.* 2005). We identify the three clusters as Island (Japan), Mainland North (Northeast China and Eastern Russia), and Mainland South (Eastern China and South Korea) (Figure 1). This clustering corresponds primarily to physical barriers to migration and accords well with previously published studies of population structure in *G. soja* (Kuroda *et al.* 2006; Kaga *et al.* 2012; Guo *et al.* 2012). Principle component analysis (PCA) of the genetic data identified a similar pattern of genetic clustering (Figure S2). The first principle component (PC) explained 5.1% of the variation and primarily separated samples on an east-west gradient. The second PC explained 2.7% of the variation and a largely north-south gradient. A Mantel test identified isolation by distance as the primary driver of population structure in our sample ($r = 0.58$, $p < 0.001$).

Allele Frequency Differentiation, Pairwise Diversity, and Linkage Disequilibrium

The Island and Mainland South clusters include the majority of accessions, 216 and 275 respectively. The Mainland North cluster was smaller with 42 individuals. *G. soja* had a mean pairwise similarity of ~70% across all samples (Figure S4A). The Mainland North population was an outlier in terms of mean percent pairwise similarity within the three clusters, a smaller number of segregating sites, and an increased number of rare variants in the folded site frequency spectrum (Table 1; Figure S4B). The Island cluster had the highest private allele richness (corrected for sample size), followed by the Mainland South, then Mainland North (Table 1). There were no fixed differences between populations. Based on the Weir and Cockerham (1984) estimator of F_{ST}, the genome-wide average single SNP $F_{ST} = 0.1$ across the entire sample. Additionally, the genome-wide average Mainland South by Mainland North $F_{ST} = 0.11$, Mainland South by Island $F_{ST} = 0.07$, and Mainland North by Island $F_{ST} = 0.18$. The average pairwise LD in the euchromatic regions, with an average half-life of D′ = 34 kb, was substantially lower than the pericentormeric regions, with an average half-life of D′ = 500 kb. This was similar to previous reported values for *G. soja* (Zhou *et al.* 2015). Curves of LD decay over distance are shown in Figure S5.

Environmental Variability and Interdependence

Environmental data was gathered from two large public databases for each of the 273 unique sampling localities at approximately 1 square km resolution. All 82 environmental variables showed a wide distribution across these sampled locations (Figure S1). Based on WorldClim records, across the range of *G. soja*, the northwestern region is colder and drier (Figure S6A and 6B). The soils data, from the ISRIC database, indicate the portion of the range in Japan has lower organic matter, higher sand content, and more variable pH than the mainland (Figure S6C). A PCA of the bioclimatic and biophysical variables generally recapitulates the geography (Figure S6D). The first four principle components explained 86.3% of the variation. Specifically, the first PC was associated with temperature, the second PC with precipitation seasonality (coefficient of variation in yearly precipitation), the third PC precipitation/soil, and the fourth PC with soil. Pearson correlations between all bioclimatic and biophysical variables showed high correlation between topsoil and subsoil (>0.99), temperature of adjacent months (>0.91), and precipitation within seasons of spring, summer, and fall. Oddly, while precipitation in July and August was highly correlated (0.86), they had low correlation with adjacent months (June-July precipitation = 0.36; August-September precipitation = 0.18).

Environmental Association Mapping

Environmental association mapping parameters were first tested with the environmental variables “Mean Temperature Wettest Quarter” and “Mean Annual Temperature” (Figure S3). Mixed models that incorporated the *K*-matrix and *Q* + *K* matrix outperformed a naïve model or Q-matrix only model (Figure S3). When comparing the genomic inflation parameter, Λ, average values were similar across variables, but the *Q* + *K* model had higher variance. Therefore, a *K*-matrix model was utilized as no additional information was gained when adding the *Q*-matrix (Figure S3). This may be because the best fit model for genetic assignment with *K* = 3 resulted in many individuals with partial assignment. Similarly, the addition of latitude as a covariate also improved Λ.

This model was applied across all 82 environmental variables. As expected with a marker set this size, typical thresholds of 0.01 Benjamini-Hochberg FDR-value or 0.001 *p*-value resulted in thousands of markers below the significance threshold (on average 1159 or 77 markers per environmental variable association). These thresholds were therefore deemed insufficient for extracting only major loci contributing to local adaptation. Instead, the threshold was set at 0.01% for each association, corresponding to the three strongest maker associations for each bioclimatic and biophysical variable. At this significance level a total of 110 unique SNPs were associated with at least one bioclimatic or biophysical variable (Table S2). We examined GO terms and the putative function of *Arabidopsis thaliana* orthologs for all genes within 34 kb (average euchromatic half-life of LD in our sample) of these significant markers. As expected, a number of patterns arose corresponding to correlated environmental variables and major contributors to the environmental PCA (Figure S6D).

Mean Temperature Wettest Quarter was a major contributor to PC1 of the environmental PCA. Mixed model association of this variable identified an association on chromosome 8 with two SNPs (*p* = 1.47E-6 and 6.78E-6) occurring less than 5 kb away from Glyma.08g298200 (Figure 2A; Figure 2B). The Arabidopsis ortholog is MYB88 (Soybase), functionally annotated as "Encodes a putative transcription factor involved in stomata development". The non-reference alleles for the two significant markers at this locus are more common in *G. soja* than in landrace and elite soybean lines (Figure 2C).

The environmental trait distribution of Mean Temperature Wettest Quarter reveals that while both the reference and non-reference alleles are found in all three population clusters (Figure S7A), individuals with the non-reference allele occur in environments that are ~2˚ C warmer in the wettest quarter than those with the reference variant, on average (Figure S7B).

Mixed model associations with monthly temperature often exhibited a pattern in which adjacent months were associated with mostly the same variants. One striking occurrence was SNP 'BARC_1.01_Gm16_1552499_A_G,' significant in 11 temperature based bioclimatic variable associations including: Max Temperature Warmest Month, Mean Temperature Warmest Quarter, Maximum Temperature June, July, and August, Mean Temperature May, June, July, and August, and Minimum Temperature June and July (Figure S8). Many of these bioclimatic variables were major contributors to PC1 (temperature) in the environmental PCA. The Arabidopsis ortholog for the nearest gene, Glyma.16g017600, was TMP14, which encodes the P subunit of Photosystem I (Khrouchtchova *et al.* 2005). Individuals with the reference variant had higher frequency at sites with cooler temperatures (Figure S9).

One of the strongest associations with monthly precipitation was between SNP 'BARC_1.01_Gm_08_2254106_G_A,' on chromosome 8, and July Precipitation (Figure S10). This SNP was significant in both July Precipitation and Precipitation Wettest Quarter. Associations with monthly precipitation overlapped less frequently with adjacent months. This was likely due to the lower correlation found between adjacent month's precipitations. This SNP falls within Glyma.08g028200 (Figure S10). The Arabidopsis orthologs for this gene is PECT1, known to be involved in respiration capacity in leaves (Otsuru *et al.* 2013).

The strongest association with a biophysical (soil) variable was an association between SNP ‘BARC_1.01_Gm14_23750665_G_A,’ on chromosome 14, and Percent Sand Subsoil (Figure 3A). This marker was also significant for Percent Sand Topsoil, Percent Silt Topsoil and Subsoil, and Cation Exchange Capacity Topsoil. As shown in Figure 3, this SNP occurs in a pericentromeric region with low SNP density in the SoySNP50K assay. The gene Glyma.14g141200 occurs in this region, and has the Arabidopsis ortholog YUC6 (Figure 3B), which is involved in the auxin biosynthesis pathway that provides enhanced resistance to water stress (Kim *et al.* 2013). The non-reference variant was rare in our sample and not present in elite lines or landraces in a previous study (Figure 3C) (Song *et al*. 2013). Distributions of Topsoil Percent Silt and Percent Sand content reveal that individuals carrying the non-reference variant were present in locations with 6% higher percent silt and 9% lower percent sand on average, than individuals carrying the reference variant (Figure 3D, Figure S11). This is not merely a result of population structure, as the accessions with the non-reference variant are widely dispersed (Figure 3E).

For a complete list of significant markers see Table S2 or Figure S12 for a Manhattan plot of the mixed model association results for all of the bioclimatic and biophysical variables individually.

F_{ST} and SPA Outliers

SPA and F_{ST} outlier analyses were used to identify allele frequency that could indicate the action of selection at a locus. SPA explored allele frequency differentiation across the geographic range. F_{ST}, an estimate of allele frequency differentiation between populations, was evaluated on a SNP by SNP basis in addition to the genome-wide averages described above. SPA outliers tended to divide the sample along the same axes identified in genetic assignment and PCA analyses. Overall, SPA selection scores were positively correlated with F_{ST} ($r = 0.76$, $r^2 = 0.58$, Figure S13 & Figure S14). The 99.9% outliers from SPA (Table S3) and F_{ST} (Table S4) overlapped at two SNPs (9%) (Figure 4A). One of these SNPs was the highest SPA selection score. These two outlier SNPs occur in a gene family cluster on chromosome 15 (Figure 4B) with Arabidopsis orthologs annotated as, “bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein” (Figure 4C). Three genes in this family were previously found to be duplicated or deleted in a sampling of modern soybean lines (Anderson *et al*. 2014) (Figure 4). Overall, SPA outliers were significantly enriched for genic space whereas F_{ST} outliers and significant environmental associations were enriched for non-genic space (Table S5).

DISCUSSION

Population Structure of Glycine soja

Genetic assignment analysis of *G. soja* in East Asia identified three primary clusters: Mainland North, Mainland South, and Island. The Mantel test identified isolation by distance as the primary contributor to this clustering, a result consistent with genetic drift (Cregan and Hartwig 1984; Nakayama and Yamaguchi 2002). The Sea of Japan likely also contributed as a physical barrier between the Island cluster and the mainland clusters. All clusters had relatively small private allele richness compared to other plant species (Fang *et al.* 2014; Cornille *et al.* 2015). Uniquely, the Mainland North cluster had a pattern of elevated Weir and Cockerham F_{ST} and the smallest number of polymorphic SNPs. Small population size and a higher level of genetic drift likely contributed to divergence in allele frequencies in this cluster. One concern is this unusual pattern in the Mainland North cluster was simply a result of ascertainment bias associated with a fixed SNP platform. Within the SoySNP50K discovery panel, two individuals were *G. soja*: PI468916 (Mainland South) and PI479752 (collected in China but does not have longitude and latitude) (Song *et al*. 2013). Therefore, both the Mainland North and Island clusters did not have a member in the discovery panel suggesting ascertainment bias is less likely to be the primary source of the elevated Weir and Cockerham F_{ST} patterns found in the Mainland North cluster.

Agronomic Implications of Environmental Association

The identification of genetic variants associated with higher temperatures or lower moisture could contribute to an understanding of the potential genetic basis of plant response to global climate change. The variant detected in association with “Mean Temperature Wettest Quarter” on chromosome 8 is associated with ~ 2˚ C higher temperature than the reference variant. The non-reference allele at this locus is found at ~ 90% of our *G. soja* samples while a previous study found this variant at 3.6% in elite soybean (Song *et al*. 2013). The Arabidopsis ortholog (MYB88) of a nearby gene is involved in stomata development and drought stress response (Xie *et al.* 2010). Another locus with potential effect on temperature response was detected on chromosome 16. The Arabidopsis ortholog for the nearest gene was TMP14, which encodes the P subunit of Photosystem I (Khrouchtchova *et al.* 2005). This variant was present in accessions sampled from the Korean Peninsula north into Russia, and corresponded to cooler growing season temperatures. The variant detected occurs at a moderate frequency (30%) in elite lines. These findings suggest a naturally occurring variant at these loci could contribute to improved drought response in elite soybeans.

The inclusion of soils data permits the exploration of environmental variables not previously explored, with the caveat that soil characteristics vary on a finer scale and thus are less readily generalizable than patterns such as temperature or rainfall regimes (Brady *et al.* 2005). We divided the data into topsoil and subsoil. While most of the root mass in soybeans occurs in the topsoil, only rarely were different markers found significant for associations in topsoil than subsoil. Soil texture and content associations did identify a number of associations that have potential agronomic applications. For example, SNP 'BARC_1.01_Gm04_3461538_T_C', on chromosome 4 was associated with "soil pH" (Figure S15) which may be relevant to response to iron deficiency chlorosis (IDC). IDC is not necessarily a shortage of iron in the soil but the inability of the plant to uptake iron under certain conditions (Hansen *et al.* 2003). IDC has a number of soil and environmental factors associated with its severity, including high early season moisture, low temperature, and high soil pH (Hansen *et al*. 2003). The gene closest to the variant (3 kb away) is Glyma.04g044000, whose reported Arabidopsis ortholog is NRAMP2, known to be essential for Arabidopsis seed germination and development in low iron conditions (Lanquar *et al.* 2005).

Utility of Complementary Approaches

The SPA and F_{ST} outlier estimates were highly correlated, but identified no overlap with the environmental association results. This was not unexpected as SPA and F_{ST} outlier loci are identified based on frequency difference in populations but are not predicated on environmental variation. Differing assumptions in outlier analysis can readily shift the most extreme outliers in the distribution from the top positions in an empirical distribution (reviewed in(Akey 2009)). Reduced recombination rates, as observed in pericentromeric regions of *G. max* (Schmutz *et al.* 2010), can contribute to elevated allele frequency divergence (as measured by F_{ST}). This effect has been attributed to the effects of linked selection (Charlesworth *et al.* 1993; Andolfatto *et al.* 1999) and has been observed in a number of crop wild progenitors, including wild barley (Fang *et al.* 2014) and teosinte (Yamasaki *et al.* 2005).

As noted in the Results, we observed the co-occurrence of an F_{ST} and SPA outlier at a locus previously reported with copy number variation (Anderson *et al*. 2014). The co-occurrence of significant allele frequency gradients in *G. soja* and copy number variation in cultivated soybean suggests the potential for the contribution of copy number variation to adaptive phenotypes in the wild. Recently, there has been increased interest in understanding the link between phenotypic variation and fine-scale structural variation: deletions or duplications ranging in size from single genes to sizeable pieces of chromosomes. New techniques have identified many important phenotypes controlled by this type of genomic variation (Żmieńko *et al.* 2014). However, it should be noted that we were unable to investigate copy number variation in *G. soja* with this dataset as we were querying single SNPs, which are not diagnostic of chromosomal structural variation, so the broader implications of the co-occurrence is unclear.

Utility to Plant Breeding

These findings in *G. soja* could be especially beneficial for plant breeders focusing on abiotic stress tolerance. Compared to other major staple crops, soybean improvement has rarely tapped into the genetic potential found in its crop wild relative (Hajjar and Hodgkin 2007). This underutilization is likely related to the amount of effort required to select lines from overwhelmingly large germplasm collections, make a multitude of crosses to create large mapping populations, and properly phenotype large populations for specific traits. Environmental association is an attractive alternative, where the large diversity in a germplasm collection is used to scan for local adaptation to specific environmental or abiotic factors. Targeted backcrossing to introgress only the putatively beneficial variant into relevant backgrounds, followed by phenotyping, would validate the loci identified. We identify such promising loci (Table S2, S3, and S4), which could immediately be applied to a validation population and shortly implemented in a breeding program.

Limitations and Biases

Local adaptation of *G. soja,* detected through SPA, F_{ST} outliers, and environmental association, can potentially provide variants linked to untapped genetic adaptations to abiotic stress tolerance. While these results are promising, a number of limitations and biases need also to be considered with these data and methods. The potential to identify putatively adaptive environmental associations is limited by the resolution at which the bioclimatic and biophysical data were collected. This is especially true of soils data, which may vary over finer scales than those at which the data were collected. It is also important to note the SNPs identified are not causative variants but rather presumed to be in LD with a causative variant. Major QTL can easily be missed due to insufficient marker coverage or simply not be in LD with the segregating markers available. There are several limitations of the association-mapping framework. The first is that differences in variant frequency are merely associated with the environmental variables measured. Individual environmental factors may not constitute the selective pressure that generates this putatively adaptive difference (Tiffin and Ross-Ibarra 2014). Second, the ability to detect associations is conditioned on the ability to detect a difference in distributions between allelic states, meaning that sample size and allele frequency are limiting factors. Next, a fixed SNP platform is being used, and thus we must make the assumption that relatively common variants contribute to adaptive variations (see (Morrell *et al.* 2011) for more on the common trait, common variants assumption). Any ascertainment bias in making this SNP platform and the finite number of markers available can affect results. This implies that it is unlikely that rare alleles of large effect will be identified as they will not be genotyped and are unlikely to be in strong LD with a queried marker (Thornton *et al.* 2013). Also, both F_{ST} outlier and association analyses have greater power to detect functional variants that are subject to antagonistic pleiotropy (i.e., those that are advantageous in one environment and deleterious in others) rather than loci that exhibit conditional neutrality (i.e., the advantageous in one environment and neutral in others) (Tiffin and Ross-Ibarra 2014) .

Conclusion

Four large public databases (GRIN, WorldClim, ISRIC, and Soybase) were used to explore the intersection of bioclimatic, biophysical, and genetic components of the important soybean crop wild relative, *G. soja*. Genetic variation associated with the environmental variation across the native range of *G. soja* was identified. While many studies have used crop wild relatives to study biotic stress (Hajjar and Hodgkin 2007), here we provide an approach aimed at identifying novel loci that could contribute to abiotic stress tolerance. The ability to identify loci associated with local adaptation to environmental variables provides an opportunity to utilize crop wild relatives in a targeted manner to address issues related to crop improvement; or issues likely to be exacerbated by a changing global climate.

G. soja has been used to explore the genetic basis of many traits such as yield, protein content, and biotic stress (Sebolt *et al.* 2000; Wang *et al.* 2001; Concibido *et al.* 2003), but has been relatively untapped in soybean improvement (Hajjar and Hodgkin 2007). This genome scan of a germplasm collection can be viewed as “population genetics enabled breeding,” the use of population genetics techniques to provide a targeted list of genomic regions for introgression and pre-breeding. The method of targeted germplasm evaluation used here could prove useful in collaboration with recent initiatives to categorize and evaluate the world’s germplasm collections (www.DivSeek.org, (McCouch *et al.* 2013; Dempewolf *et al.* 2014)). Ideally these results can play a role in improving crop tolerance to our globally changing abiotic conditions.

Table 1. Diversity summary statistics within assigned clusters of *Glycine soja* sampled.

Population	**Sample Size**	**Segregating sites**	**Private allelic richness**	**Percent pairwise difference**
Island	216	31,698	0.025 (0.011)	0.340
Mainland South	275	32,360	0.009 (0.005)	0.337
Mainland North	42	23,797	0.001 (0.0001)	0.306
Mainland South + Island	492	32,416	0.25 (0.16)	0.349
Mainland North + Island	258	32,350	0.006 (0.002)	0.345
Mainland South + Mainland North	317	32,360	0.045 (0.029)	0.338

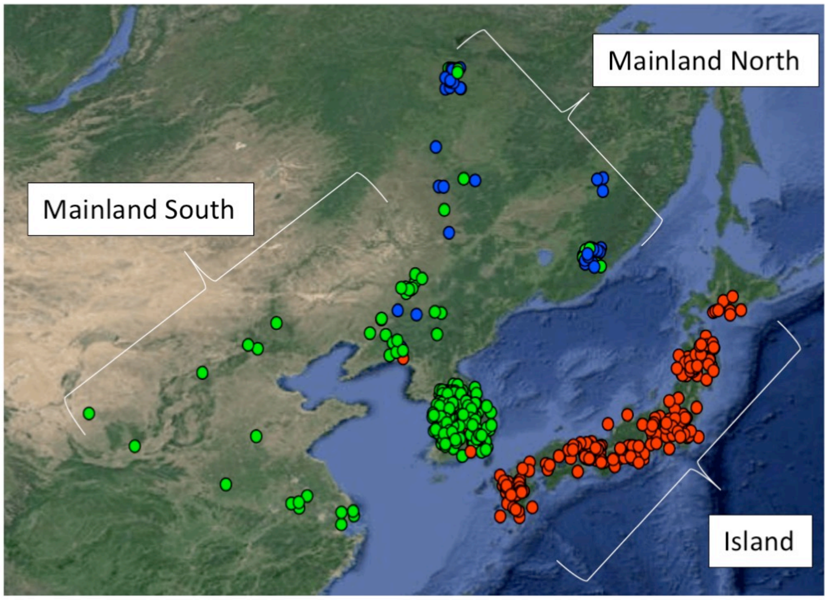
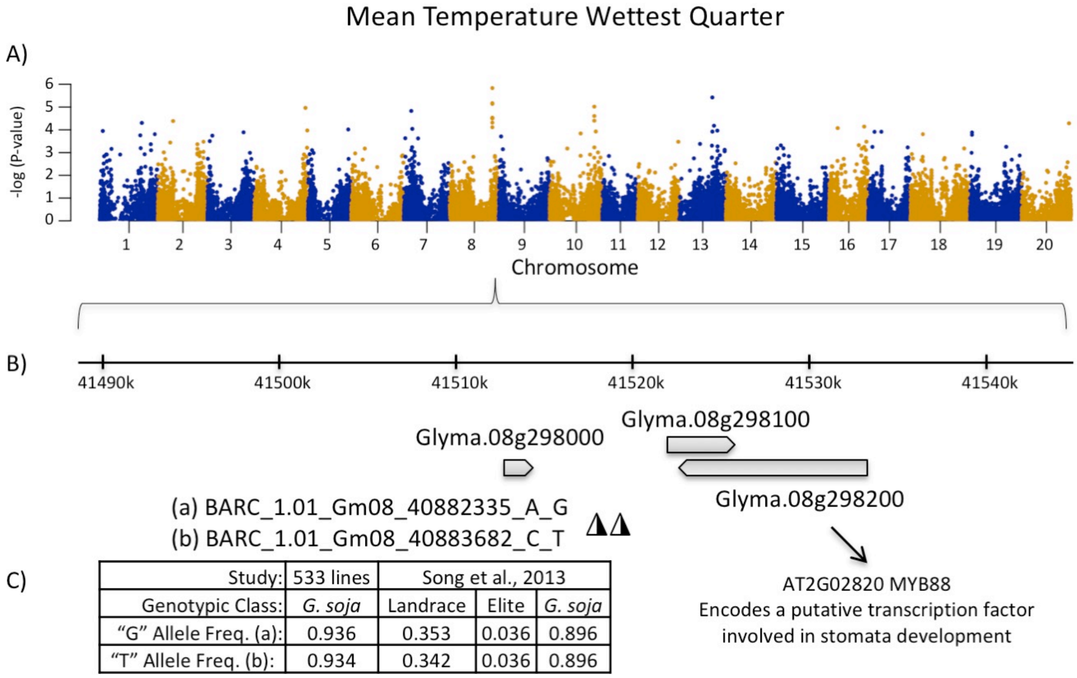
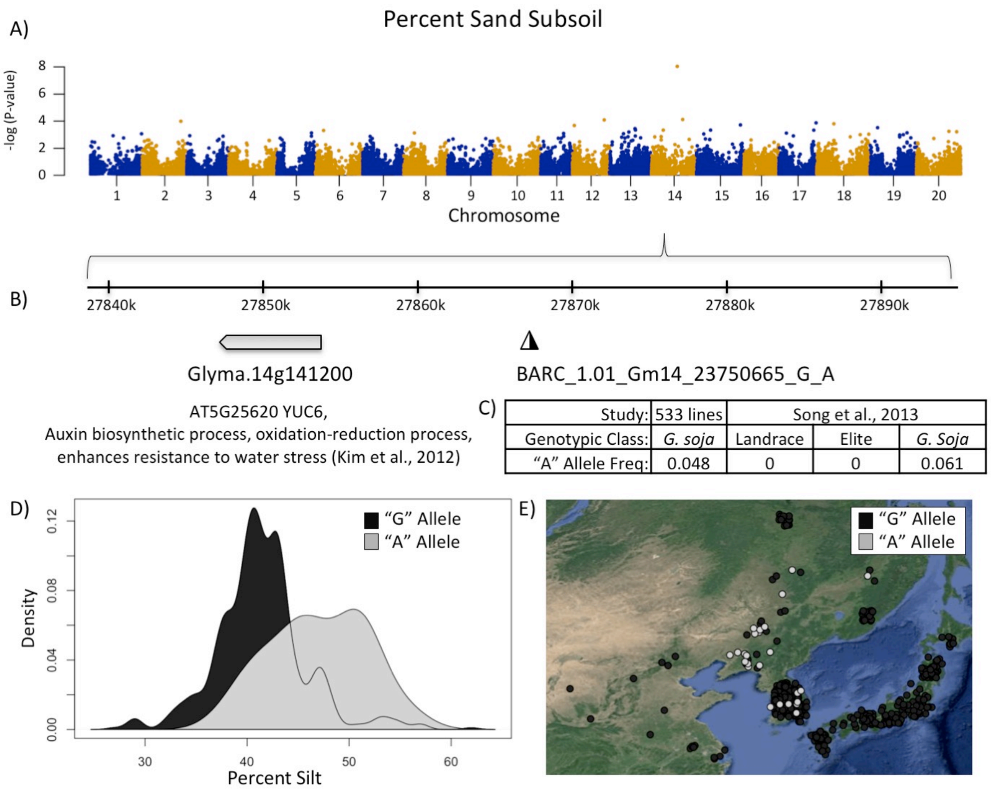


Figure 1. Results of STRUCTURE analysis in *G. soja* accessions and the geographical location in which each were collected. The spot colors correspond to the STRUCTURE assignment of each accession, Green: Mainland South; Blue: Mainland North; Red: Island. The assignment of samples into three genetic clusters generally accords with geography. The spots have been jittered to show overlapping samples.



Study:	533 lines	Song et al., 2013		
Genotypic Class:	*G. soja*	Landrace	Elite	*G. soja*
"G" Allele Freq. (a):	0.936	0.353	0.036	0.896
"T" Allele Freq. (b):	0.934	0.342	0.036	0.896

Figure 2. Genome-wide associations with Mean Temperature Wettest Quarter. A) Manhattan plot of negative log p-values. B) Zoom in on 60 kb region around the significant markers BARC_1.01_Gm08_40882335_A_G and BARC_1.01_Gm08_40883682_C_T. The Arabidopsis homolog for a near gene, Glyma.08g298200, is MYB88, a gene associated stomata development. C) The frequency of non-reference "G" and "T" alleles is high in G. soja and rare in a previous study of landrace and elite lines.



Study:	533 lines	Song et al., 2013		
Genotypic Class:	*G. soja*	Landrace	Elite	*G. Soja*
"A" Allele Freq:	0.048	0	0	0.061

Figure 3. Genome-wide association results of percent sand and percent silt. A) Genome wide view of association results for percent sand topsoil. B) Zoom in on 60 kb region around the significant marker BARC_1.01_Gm14_23750665_G_A, the most significant hit for topsoil and subsoil percent sand, and topsoil and subsoil percent silt. The "A" allele at this locus is associated with high silt environments and is not found in a previous scan of landrace and elite soybean cultivars. The Arabidopsis best hit for the nearest gene, Glyma.14g141200, is YUC6, a gene associated with enhanced resistance to water stress. C) The "A" allele is rare in our sample and found to be rare or not present in a previous screen of soybean genotypic classes (Song *et al.* 2013). D) Density plot of allele frequency distribution for Percent Silt. The individuals with the "G" allele are shaded in dark gray overlaid with the "A" allele individuals in light gray. E) Geographic location of individuals with the "G" allele (Dark gray) or "A" allele (light gray) with jitter added to show overlapping samples. Individuals with missing genotyping data at this SNP are not shown.

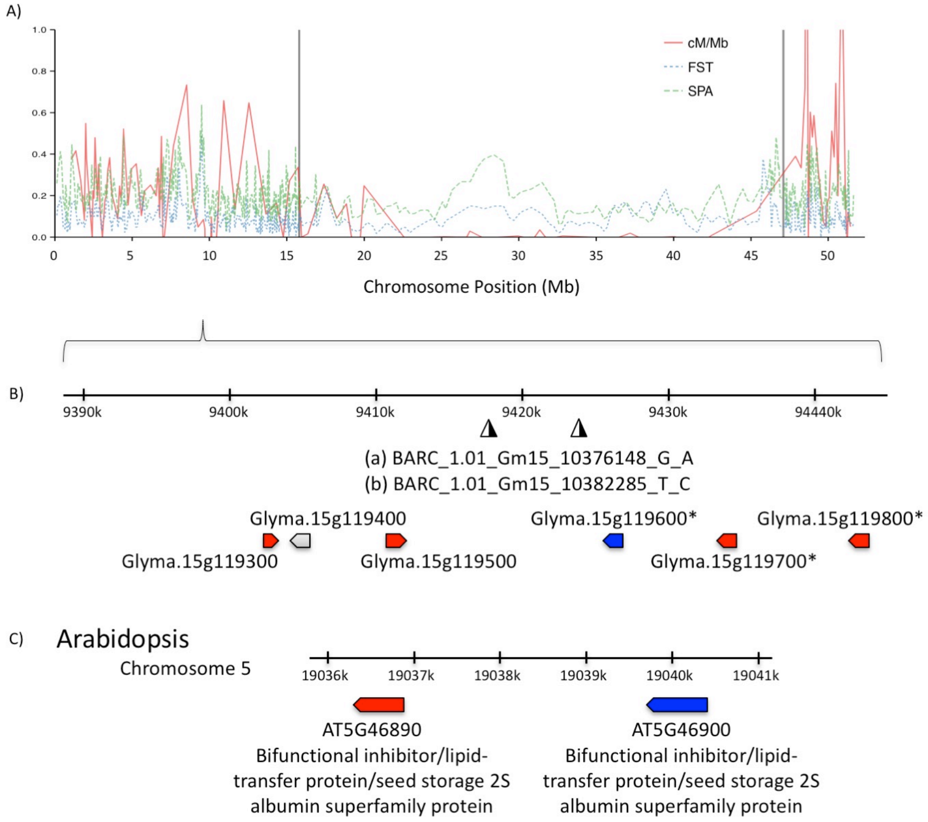


Figure 4. SPA, FST, and recombination rate in the *G. soja* genome. A) Sliding window of these values plotted on chromosome 15. Recombination decreases dramatically through the pericentromeric region, denoted by the the vertical gray dotted lines. B) Zoom in on 60 kb region around significant SPA markers BARC_1.01_Gm15_10376148_G_A and BARC_1.01_Gm15_10382285_T_C. A region of notably low recombination and both high FST and SPA values. Three genes in this region (denoted with asterisks) were previously found to be duplicated or deleted in elite soybean lines (Anderson *et al.* 2014). This cluster of genes appear to be members of a gene family. The Arabidopsis best hit for the genes denoted in red is AT5G46890, a Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein. Similarly, The Arabidopsis top hit for Glyma.15g119600, denoted in blue, is AT5G46900, a Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein. The implications of structural variation relating to FST, SPA hits, or recombination are not yet clear.

References

Abecasis, G. R., E. Noguchi, A. Heinzmann, J. A. Traherne, S. Bhattacharyya *et al.*, 2001 Extent and distribution of linkage disequilibrium in three genomic regions. Am. J. Hum. Genet. 68: 191–197.

Ahmad, Q. N., E. J. Britten, and D. E. Byth, 1979 Inversion heterozygosity in the hybrid soybean x Glycine soja: Evidence from a pachytene loop configuration and other meiotic irregularities. J. Hered. 70: 358–364.

Akey, J. M., 2009 Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 19: 711–22.

Anderson, J. E., M. B. Kantar, T. Y. Kono, F. Fu, A. O. Stec *et al.*, 2014 A roadmap for functional structural variants in the soybean genome. G3 4: 1307–1318.

Andolfatto, P., J. D. Wall, and M. Kreitman, 1999 Unusual Haplotype Structure at the Proximal Breakpoint of In(2L)t in a Natural Population of Drosophila melanogaster. Genetics 153: 1297–1311.

Andrews, S., 2010 FastQC: A quality control tool for high throughput sequence data.

Ashfield, T., A. Bocian, D. Held, A. D. Henk, L. F. Marek *et al.*, 2007 Genetic and Physical Localization of the Soybean Rpg1-b Disease Resistance Gene Reveals a Complex Locus Containing Several Tightly Linked Families of NBS-LRR Genes.

Beaumont, M. A., 2005 Adaptation and speciation: what can Fst tell us? Trends Ecol. Evol. 20: 435–440.

Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. 13: 969–980.

Belfield, E. J., X. Gan, A. Mithani, C. Brown, C. Jiang *et al.*, 2012 Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of Arabidopsis thaliana. Genome Res. 22: 1306–1315.

Bernard, R. L., and C. R. Cremeens, 1988 Registration of “Williams 82” soybean. Crop Sci. 28: 1027–1028.

Berriz, G. F., J. E. Beaver, C. Cenik, M. Tasan, and F. P. Roth, 2009 Next generation software for functional trend analysis. Bioinformatics 25: 3043–3044.

Bikard, D., D. Patel, C. Le Metté, V. Giorgi, C. Camilleri *et al.*, 2009 Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science 323: 623–6.

Birchler, J. A., and R. A. Veitia, 2012 Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc. Natl. Acad. Sci. U. S. A. 109: 14746–14753.

Bolon, Y.-T., A. O. Stec, J.-M. Michno, J. Roessler, P. B. Bhaskar *et al.*, 2014 Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. Genetics 198: 967–981.

Bolon, Y.-T., W. Haun, W. Xu, D. Grant, M. Stacey *et al.*, 2011 Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. Plant Physiol. 156: 240–253.

Bomblies, K., and D. Weigel, 2007 Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nat. Rev. Genet. 8: 382–393.

Boocock, J., D. Chagné, T. R. Merriman, and M. A. Black, 2015 The distribution and impact of common copy-number variation in the genome of the domesticated apple, Malus x domestica Borkh. BMC Genomics 16: 848.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–5.

Brady, K. U., A. R. Kruckeberg, and H. D. Bradshaw Jr., 2005 Evolutionary Ecology of Plant Adaptation to Serpentine Soils. Annu. Rev. Ecol. Evol. Syst. 36: 243–266.

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. 43: 956–963.

Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Chen, K., J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki *et al.*, 2009 BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6: 677–81.

Cheng, K. C., J. Beaulieu, E. Iquira, F. J. Belzile, M. G. Fortin *et al.*, 2008 Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. J. Agric. Food Chem. 56: 3057–3067.

Chia, J.-M., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44: 803–807.

Chung, G., and R. J. Singh, 2008 Broadening the Genetic Base of Soybean: A Multidisciplinary Approach. CRC. Crit. Rev. Plant Sci. 27: 295–341.

Chung, W. H., N. Jeong, J. Kim, W. K. Lee, Y. G. Lee *et al.*, 2014 Population structure and domestication revealed by high-depth resequencing of korean cultivated and wild soybean genomes. DNA Res. 21: 153–167.

Clark, K. A., and P. J. Krysan, 2010 Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines. Plant J. 64: 990–1001.

Concibido, V. C., B. La Vallee, P. McLaird, N. Pineda, J. Meyer *et al.*, 2003

Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. Theor. Appl. Genet. 106: 575–82.

Cook, D. E., A. M. Bayless, K. Wang, X. Guo, Q. Song *et al.*, 2014 Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode. Plant Physiol. 165: 630–647.

Cook, D. E., T. G. Lee, X. Guo, S. Melito, K. Wang *et al.*, 2012 Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean. Science (80-.). 338: 1206–1209.

Cornille, A., A. Feurtey, U. Gélin, J. Ropars, K. Misvanderbrugge *et al.*, 2015 Anthropogenic and natural drivers of gene flow in a temperate wild fruit tree: a basis for conservation and breeding programs in apples. Evol. Appl. 8: 373–84.

Cregan, P. B., and E. E. Hartwig, 1984 Characterization of Flowering Response to Photoperiod in Diverse Soybean Genotypes1. Crop Sci. 24: 659.

Curtin, S. J., F. Zhang, J. D. Sander, W. J. Haun, C. Starker *et al.*, 2011 Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. Plant Physiol. 156: 466–473.

Dangl, J. L., and J. D. Jones, 2001 Plant pathogens and integrated defence responses to infection. Nature 411: 826–33.

Dempewolf, H., R. J. Eastwood, L. Guarino, C. K. Khoury, J. V. Müller *et al.*, 2014 Adapting Agriculture to Climate Change: A Global Initiative to Collect, Conserve, and Use Crop Wild Relatives. Agroecol. Sustain. Food Syst. 38: 369–377.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Díaz, A., M. Zikhali, A. S. Turner, P. Isaac, and D. A. Laurie, 2012 Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). PLoS One 7: e33234.

Du, J., Z. Tian, Y. Sui, M. Zhao, Q. Song *et al.*, 2012 Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. Plant Cell 24: 21–32.

Dunham, M. J., H. Badrane, T. Ferea, J. Adams, P. O. Brown *et al.*, 2002 Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. 99: 16144–9.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra *et al.*, 2010 Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics 185: 969–82.

Endo, M., M. Kumagai, R. Motoyama, H. Sasaki-Yamagata, S. Mori-Hosokawa *et al.*, 2014 Whole-Genome Analysis of Herbicide-Tolerant Mutant Rice Generated by Agrobacterium-Mediated Gene Targeting. Plant Cell Physiol. 56: 116–125.

Epstein, B., M. J. Sadowsky, and P. Tiffin, 2014 Selection on horizontally transferred and duplicated genes in sinorhizobium (ensifer), the root-nodule symbionts of medicago. Genome Biol. Evol. 6: 1199–209.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14: 2611–20.

Fang, Z., A. M. Gonzales, M. T. Clegg, K. P. Smith, G. J. Muehlbauer *et al.*, 2014 Two Genomic Regions Contribute Disproportionately to Geographic Differentiation in Wild Barley. G3 4: 1193–1203.

Fang, Z., T. Pyhajarvi, A. L. Weber, R. K. Dawe, J. C. Glaubitz *et al.*, 2012 Megabase-scale inversion polymorphism in the wild ancestor of maize. Genetics 191: 883–894.

Findley, S. D., A. L. Pappas, Y. Cui, J. A. Birchler, R. G. Palmer *et al*., 2011 Fluorescence in situ hybridization-based karyotyping of soybean translocation lines. G3 (Bethesda). 1: 117–29.

Findley, S. D., S. Cannon, K. Varala, J. Du, J. Ma et al., 2010 A fluorescence in situ hybridization system for karyotyping soybean. Genetics 185: 727–44.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger *et al.*, 2010 The Pfam protein families database. Nucleic Acids Res. 38: D211–22.

Flor, H. H., 1971 Current Status of the Gene-For-Gene Concept. Annu. Rev. Phytopathol. 9: 275–296.

Forsbach, A., D. Schubert, B. Lechtenberg, M. Gils, and R. Schmidt, 2003 A comprehensive characterization of single-copy T-DNA insertions in the Arabidopsis thaliana genome. Plant Mol. Biol. 52: 1–16.

Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt *et al.*, 2011 A map of local adaptation in Arabidopsis thaliana. Science 334: 86–9.

Gaines, T. A., D. L. Shaner, S. M. Ward, J. E. Leach, C. Preston *et al.*, 2011 Mechanism of resistance of evolved glyphosate-resistant Palmer amaranth (Amaranthus palmeri). J. Agric. Food Chem. 59: 5886–5889.

Gaines, T. A., W. Zhang, D. Wang, B. Bukun, S. T. Chisholm *et al.*, 2010 Gene amplification confers glyphosate resistance in Amaranthus palmeri. Proc. Natl. Acad. Sci. U. S. A. 107: 1029–1034.

Garsmeur, O., J. C. Schnable, A. Almeida, C. Jourda, A. D’Hont *et al.*, 2014 Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31: 448–454.

Gill, N., S. Findley, J. G. Walling, C. Hans, J. Ma *et al.*, 2009 Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 151: 1167–1174.

Gordon, S. G., S. K. St. Martin, and A. E. Dorrance, 2006 Rps8 Maps to a Resistance Gene Rich Region on Soybean Molecular Linkage Group F. Crop Sci. 46: 168.

Grant, D., R. T. Nelson, S. B. Cannon, and R. C. Shoemaker, 2010 SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38: D843–6.

Gresham, D., M. J. Dunham, and D. Botstein, 2008 Comparing whole genomes using DNA microarrays. Nat. Rev. Genet. 9: 291–302.

Gu, W., F. Zhang, and J. R. Lupski, 2008 Mechanisms for human genomic rearrangements. Pathogenetics 1: 4.

Guo, J., Y. Liu, Y. Wang, J. Chen, Y. Li *et al.*, 2012 Population structure of the wild soybean (Glycine soja) in China: implications from microsatellite analyses. Ann. Bot. 110: 777–85.

Hajjar, R., and T. Hodgkin, 2007 The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156: 1–13.

Han, J. J., D. Jackson, and R. Martienssen, 2012 Pod corn is caused by rearrangement at the Tunicate1 locus. Plant Cell 24: 2733–2744.

Hancock, C. N., F. Zhang, K. Floyd, A. O. Richardson, P. Lafayette *et al.*, 2011 The rice miniature inverted repeat transposable element mPing is an effective insertional mutagen in soybean. Plant Physiol. 157: 552–562.

Hansen, N. C., M. A. Schmitt, J. E. Anderson, and J. S. Strock, 2003 Iron Deficiency of Soybean in the Upper Midwest and Associated Soil Properties. Agron. J. 95: 1595.

Harlan, J. R., J. M. J. de Wet, and E. G. Price, 1973 Comparative Evolution of Cereals. Evolution (N. Y). 27: 311–325.

Hastings, P. J., J. R. Lupski, S. M. Rosenberg, and G. Ira, 2009 Mechanisms of change in gene copy number. Nat. Rev. Genet. 10: 551–64.

Haun, W. J., D. L. Hyten, W. W. Xu, D. J. Gerhardt, T. J. Albert *et al.*, 2011 The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol. 155: 645–655.

Haun, W., A. Coffman, B. M. Clasen, Z. L. Demorest, A. Lowy *et al.*, 2014 Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. Plant Biotechnol. J. 12: 934–940.

Hayes, A. J., S. C. Jeong, M. A. Gore, Y. G. Yu, G. R. Buss *et al.*, 2004 Recombination within a nucleotide-binding-site/leucine-rich-repeat gene cluster produces new variants conditioning resistance to soybean mosaic virus in soybeans. Genetics 166:

493–503.

Heinz, D. J., and G. W. P. Mee, 1971 Morphologic, Cytogenetic, and Enzymatic Variation in Saccharum Species Hybrid Clones Derived from Callus Tissue. Am. J. Bot. 58: 257–262.

Hengl, T., J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. M. Heuvelink *et al.*, 2014 SoilGrids1km--global soil information based on automated mapping. PLoS One 9: e105992.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, 2005 Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25: 1965–1978.

Hirsch, C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni *et al.*, 2014 Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26: 121–35.

Huddleston, J., S. Ranade, M. Malig, F. Antonacci, M. Chaisson *et al.*, 2014 Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 24: 688–96.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

Hwang, W. J., M. Y. Kim, Y. J. Kang, S. Shim, M. G. Stacey *et al.*, 2014 Genome-wide analysis of mutations in a dwarf soybean mutant induced by fast neutron bombardment. Euphytica 203: 399–408.

Hyten, D. L., Q. Song, Y. Zhu, I.-Y. Y. Choi, R. L. Nelson *et al.*, 2006 Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. U. S. A. 103: 16666–71.

Hyten, D. L., S. B. Cannon, Q. Song, N. Weeks, E. W. Fickus *et al.*, 2010 High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11: 38.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe *et al.*, 2004 Detection of large-scale variation in the human genome. Nat. Genet. 36: 949–51.

Iovene, M., T. Zhang, Q. Lou, C. R. Buell, and J. Jiang, 2013 Copy number variation in potato–an asexually propagated autotetraploid species. Plant J. 75: 80–89.

Jacobs, T. B., P. R. LaFayette, R. J. Schmitz, and W. A. Parrott, 2015 Targeted genome modifications in soybean with CRISPR/Cas9. BMC Biotechnol. 15: 16.

Jain, S. M., 2001 Tissue culture-derived variation in crop improvement. Euphytica 118:

153–166.

Jiang, C., A. Mithani, X. Gan, E. J. Belfield, J. P. Klingler *et al.*, 2011 Regenerant arabidopsis lineages display a distinct genome-wide spectrum of mutations conferring variant phenotypes. Curr. Biol. 21: 1385–1390.

Jones, J. D. G., and J. L. Dangl, 2006 The plant immune system. Nature 444: 323–9.

Kaga, A., T. Shimizu, S. Watanabe, Y. Tsubokura, Y. Katayose *et al.*, 2012 Evaluation of soybean germplasm conserved in NIAS genebank and development of mini core collections. Breed. Sci. 61: 566–92.

Kanizay, L. B., T. B. Jacobs, K. Gillespie, J. A. Newsome, B. N. Spaid *et al.*, 2015 HtStuf: High-Throughput Sequencing to Locate Unknown DNA Junction Fragments. Plant Genome 8: 1–10.

Katju, V., and U. Bergthorsson, 2013 Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front. Genet. 4: 273.

Khoury, C. K., B. Heider, N. P. Castañeda-Álvarez, H. A. Achicanoy, C. C. Sosa *et al.*, 2015 Distributions, ex situ conservation priorities, and genetic resource potential of crop wild relatives of sweetpotato [Ipomoea batatas (L.) Lam., I. series Batatas]. Front. Plant Sci. 6: 251.

Khrouchtchova, A., M. Hansson, V. Paakkarinen, J. P. Vainonen, S. Zhang *et al.*, 2005 A previously found thylakoid membrane protein of 14kDa (TMP14) is a novel subunit of plant photosystem I and is designated PSI-P. FEBS Lett. 579: 4808–12.

Kim, J. I., D. Baek, H. C. Park, H. J. Chun, D.-H. Oh *et al.*, 2013 Overexpression of Arabidopsis YUCCA6 in potato results in high-auxin developmental phenotypes and enhanced resistance to water deficit. Mol. Plant 6: 337–49.

Kim, K.-S., C. B. Hill, G. L. Hartman, D. L. Hyten, M. E. Hudson *et al.*, 2010a Fine mapping of the soybean aphid-resistance gene Rag2 in soybean PI 200538. Theor. Appl. Genet. 121: 599–610.

Kim, M. Y., S. Lee, K. Van, T.-H. Kim, S.-C. Jeong *et al.*, 2010b Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. U. S. A. 107: 22032–7.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Kirkpatrick, M., 2010 How and why chromosome inversions evolve. PLoS Biol. 8: e1000501.

Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and speciation. Genetics 173: 419–34.

Klein, B. A., E. L. Tenorio, D. W. Lazinski, A. Camilli, M. J. Duncan *et al.*, 2012

Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. BMC Genomics 13: 578.

Knox, A. K., T. Dhillon, H. Cheng, A. Tondelli, N. Pecchioni *et al.*, 2010 CBF gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. Theor. Appl. Genet. 121: 21–35.

Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose, 2015 Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. 15: 1179–91.

Kovalic, D., C. Garnaat, L. Guo, Y. Yan, J. Groat *et al.*, 2012 The Use of Next Generation Sequencing and Junction Sequence Analysis Bioinformatics to Achieve Molecular Characterization of Crops Improved Through Modern Biotechnology. Plant Genome J. 5: 149–163.

Kruijt, M., M. J. D. DE Kock, and P. J. G. M. de Wit, 2005 Receptor-like proteins involved in plant disease resistance. Mol. Plant Pathol. 6: 85–97.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne *et al.*, 2009 Circos: an information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

Kuroda, Y., A. Kaga, N. Tomooka, and D. A. Vaughan, 2006 Population genetic structure of Japanese wild soybean (Glycine soja) based on microsatellite variation. Mol. Ecol. 15: 959–74.

Kyndt, T., D. Quispe, H. Zhai, R. Jarret, M. Ghislain *et al.*, 2015 The genome of cultivated sweet potato contains Agrobacterium T-DNAs with expressed genes: An example of a naturally transgenic food crop. Proc. Natl. Acad. Sci. 112: 201419685.

Ladics, G. S., A. Bartholomaeus, P. Bregitzer, N. G. Doerrer, A. Gray *et al.*, 2015 Genetic basis and detection of unintended effects in genetically modified crop plants. Transgenic Res. 24: 587–603.

Lam, H.-M., X. Xu, X. Liu, W. Chen, G. Yang *et al.*, 2010 Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42: 1053–1059.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Lanquar, V., F. Lelièvre, S. Bolte, C. Hamès, C. Alcon *et al.*, 2005 Mobilization of vacuolar iron by AtNRAMP3 and AtNRAMP4 is essential for seed germination on low iron. EMBO J. 24: 4041–51.

Latham, J. R., A. K. Wilson, and R. A. Steinbrecher, 2006 The mutational consequences of plant transformation. J. Biomed. Biotechnol. 2006: 1–7.

Lee, S., K. R. Freewalt, L. K. McHale, Q. Song, T.-H. Jun *et al.*, 2015 A high-resolution genetic linkage map of soybean based on 357 recombinant inbred lines genotyped with BARCSoySNP6K. Mol. Breed. 35: 58.

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

Lewontin, R. C., 1964 The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49: 49–67.

Lewontin, R. C., and J. Krakauer, 1973 Distribution Of Gene Frequency As A Test Of The Theory Of The Selective Neutrality Of Polymorphisms. Genetics 74: 175–195.

Li, H., and R. Durbin, 2009a Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., and R. Durbin, 2009b Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–9.

Li, Y.-H., W. Li, C. Zhang, L. Yang, R.-Z. Chang *et al.*, 2010 Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytol. 188: 242–53.

Li, Y., G. Li, J. Zhou, W. Ma, L. Jiang *et al.*, 2014 De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32: 1045–1052.

Li, Y., J. Xiao, J. Wu, J. Duan, Y. Liu *et al.*, 2012 A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. New Phytol. 196: 282–291.

Li, Y., S. Zhao, J. Ma, D. Li, L. Yan *et al.*, 2013 Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14: 579.

Li, Z., Z.-B. Liu, A. Xing, B. P. Moon, J. P. Koellhoffer *et al.*, 2015 Cas9-guide RNA Directed Genome Editing in Soybean. Plant Physiol. 169: pp.00783.2015.

Lipka, A. E., C. B. Kandianis, M. E. Hudson, J. Yu, J. Drnevich *et al.*, 2015 From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr. Opin. Plant Biol. 24: 110–8.

Lisch, D., 2013 How important are transposons for plant evolution? Nat. Rev. Genet. 14: 49–61.

Liu, B., A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada *et al.*, 2008 Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics 180: 995–1007.

Lynch, M., and A. Force, 2000 The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics 154: 459–473.

Lynch, M., and J. S. Conery, 2000 The Evolutionary Fate and Consequences of Duplicate Genes. Science (80-.). 290: 1151–1155.

Mahama, A. A., 1999 Cytogenetic analysis of translocations in Soybean. J. Hered. 90: 648–653.

Majhi, B. B., J. M. Shah, and K. Veluthambi, 2014 A novel T-DNA integration in rice involving two interchromosomal translocations. Plant Cell Rep. 33: 929–944.

Makino, T., A. McLysaght, and M. Kawata, 2013 Genome-wide deserts for copy number variation in vertebrates. Nat. Commun. 4.:

Maron, L. G., C. T. Guimaraes, M. Kirst, P. S. Albert, J. A. Birchler *et al.*, 2013 Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. U. S. A. 110: 5241–5246.

Marroni, F., S. Pinosio, and M. Morgante, 2014 Structural variation and genome complexity: is dispensable really dispensable? Curr. Opin. Plant Biol. 18: 31–6.

McClintock, B., 1931 Cytological observations of deficiencies involving known genes, translocations and an inversion in Zea mays. Missouri Agric. Exp. Stn. Res. Bull. 163: 1–30.

McCouch, S., G. J. Baute, J. Bradeen, P. Bramel, P. K. Bretting *et al.*, 2013 Agriculture: Feeding the future. Nature 499: 23–4.

McHale, L. K., W. J. Haun, W. W. Xu, P. B. Bhaskar, J. E. Anderson *et al.*, 2012 Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol. 159: 1295–1308.

McHale, L., X. Tan, P. Koehl, and R. W. Michelmore, 2006 Plant NBS-LRR proteins: adaptable guards. Genome Biol 7: 212.

McVey, M., and S. E. Lee, 2008 MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 24: 529–538.

Men, A. E., T. S. Laniya, I. R. Searle, I. Iturbe-Ormaetxe, I. Gresshoff *et al.*, 2002 Fast Neutron Mutagenesis of Soybean (Glycine soja L.) Produces a Supernodulating Mutant Containing a Large Deletion in Linkage Group H. Genome Lett. 1: 147–155.

Meyer, J. D. F., D. C. G. Silva, C. Yang, K. F. Pedley, C. Zhang *et al.*, 2009 Identification and analyses of candidate genes for rpp4-mediated resistance to Asian soybean rust in soybean. Plant Physiol. 150: 295–307.

Michelmore, R. W., and B. C. Meyers, 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 8: 1113–1130.

Michno, J.-M., X. Wang, J. Liu, S. J. Curtin, T. J. Y. Kono *et al.*, 2015 CRISPR/Cas mutagenesis of soybean and Medicago truncatula using a new web-tool and a modified Cas9 enzyme. GM Crops Food.

Miyao, A., M. Nakagome, T. Ohnuma, H. Yamagata, H. Kanamori *et al.*, 2012 Molecular spectrum of somaclonal variation in regenerated rice revealed by whole-genome sequencing. Plant Cell Physiol. 53: 256–264.

Morgante, M., E. De Paoli, and S. Radovic, 2007 Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. 10: 149–55.

Morrell, P. L., E. S. Buckler, and J. Ross-Ibarra, 2011 Crop genomics: advances and applications. Nat. Rev. Genet. 13: 85–96.

Munoz-Amatriain, M., S. R. Eichten, T. Wicker, T. A. Richmond, M. Mascher *et al.*, 2013 Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. 14: R58.

Murgia, I., D. Tarantino, C. Soave, and P. Morandini, 2011 Arabidopsis CYP82C4 expression is dependent on Fe availability and circadian rhythm, and correlates with genes involved in the early Fe deficiency response. J. Plant Physiol. 9: 168.

Muskens, M. W. M., A. P. A. Vissers, J. N. M. Mol, and J. M. Kooter, 2000 Role of inverted DNA repeats in transcriptional and post-transcriptional gene silencing. Plant Mol. Biol. 43: 243–260.

Nacry, P., C. Camilleri, B. Courtial, M. Caboche, and D. Bouchez, 1998 Major Chromosomal Rearrangements Induced by T-DNA Transformation in Arabidopsis. Genetics 149: 641–650.

Nakayama, Y., and H. Yamaguchi, 2002 Natural hybridization in wild soybean (Glycine max ssp. soja) by pollen flow from cultivated soybean (Glycine max ssp. max) in a designed population. Weed Biol. Manag. 2: 25–30.

Neelakandan, A. K., and K. Wang, 2012 Recent progress in the understanding of tissue culture-induced genome level changes in plants and potential applications. Plant Cell Rep. 31: 597–620.

Nei, M., and T. Maruyama, 1975 Letters to the editors: Lewontin-Krakauer test for neutral genes. Genetics 80: 395.

Ohno, S., 1970 Evolution by Gene Duplication. 160.

Olhoft, P. M., L. E. Flagel, and D. A. Somers, 2004 T-DNA locus structure in a large population of soybean plants transformed using the Agrobacterium-mediated cotyledonary-node method. Plant Biotechnol. J. 2: 289–300.

Orf, J. H., and B. W. Kennedy, 1992 Registration of "Bert" Soybean. Crop Sci. 32: 830.

Orf, J. H., and R. L. Denny, 2004 Registration of "MN1302" Soybean. Crop Sci. 44: 693.

Ossowski, S., K. Schneeberger, J. I. Lucas-Lledo, N. Warthmann, R. M. Clark *et al.*, 2010 The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92–94.

Otsuru, M., Y. Yu, J. Mizoi, M. Kawamoto-Fujioka, J. Wang *et al.*, 2013 Mitochondrial phosphatidylethanolamine level modulates Cyt c oxidase activity to maintain respiration capacity in Arabidopsis thaliana rosette leaves. Plant Cell Physiol. 54: 1612–9.

Palmer, R. G., H. Sun, and L. M. Zhao, 2000 Genetics and Cytology of Chromosome Inversions in Soybean Germplasm. Crop Sci. 40: 683.

Paz, M. M., J. C. Martinez, A. B. Kalvig, T. M. Fonger, and K. Wang, 2006 Improved cotyledonary node method using an alternative explant derived from mature seed for efficient Agrobacterium-mediated soybean transformation. Plant Cell Rep. 25: 206–213.

Pearce, S., R. Saville, S. P. Vaughan, P. M. Chandler, E. P. Wilhelm *et al.*, 2011 Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. Plant Physiol. 157: 1820–1831.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945–959.

Qi, Y., X. Li, Y. Zhang, C. G. Starker, N. J. Baltes *et al.*, 2013 Targeted deletion and inversion of tandemly arrayed genes in Arabidopsis thaliana using zinc finger nucleases. G3 3: 1707–15.

Qiu, J., Y. Wang, S. Wu, Y.-Y. Wang, C.-Y. Ye *et al.*, 2014 Genome re-sequencing of semi-wild soybean reveals a complex Soja population structure and deep introgression. PLoS One 9: e108479.

Qutob, D., J. Tedman-Jones, S. Dong, K. Kuflu, H. Pham *et al.*, 2009 Copy number variation and transcriptional polymorphisms of Phytophthora sojae RXLR effector genes Avr1a and Avr3a. PLoS One 4: e5066.

R Development Core Team, 2011 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ravensdale, M., A. Nemri, P. H. Thrall, J. G. Ellis, and P. N. Dodds, 2011 Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. Mol. Plant Pathol. 12: 93–102.

Robertson, A., 1975 Gene frequency distributions as a test of selective neutrality. Genetics 81: 775–85.

Rodin, S. N., and A. D. Riggs, 2003 Epigenetic silencing may aid evolution by gene duplication. J. Mol. Evol. 56: 718–29.

Roulin, A., P. L. Auer, M. Libault, J. Schlueter, A. Farmer *et al.*, 2013 The fate of duplicated genes in a polyploid plant genome. Plant J. 73: 143–53.

Sabot, F., N. Picault, M. El-Baidouri, C. Llauro, C. Chaparro *et al.*, 2011 Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant J. 66: 241–246.

Saika, H., J. Horita, F. Taguchi-Shiobara, S. Nonaka, N. Y. Ayako *et al.*, 2014 A novel rice cytochrome P450 gene, CYP72A31, confers tolerance to acetolactate synthase-inhibiting herbicides in rice and Arabidopsis. Plant Physiol. 166: 1232–1240.

Santuari, L., S. Pradervand, A.-M. Amiguet-Vercher, J. Thomas, E. Dorcey *et al.*, 2010 Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. Genome Biol. 11: 1–8.

Schatz, M. C., L. G. Maron, J. C. Stein, A. Hernandez Wences, J. Gurtowski *et al.*, 2014 Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome Biol. 15: 506.

Schmutz, J., S. B. Cannon, J. Schlueter, J. Ma, T. Mitros *et al.*, 2010 Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183.

Schnell, J., M. Steele, J. Bean, M. Neuspiel, C. Girard *et al.*, 2015 A comparative analysis of insertional effects in genetically engineered plants: considerations for pre-market assessments. Transgenic Res. 24: 1–17.

Schuler, M. A., and D. Werck-Reichhart, 2003 Functional genomics of P450s. Annu. Rev. Plant Biol. 54: 629–667.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young *et al.*, 2004 Large-scale copy number polymorphism in the human genome. Science 305: 525–8.

Sebolt, A. M., R. C. Shoemaker, and B. W. Diers, 2000 Analysis of a Quantitative Trait Locus Allele from Wild Soybean That Increases Seed Protein Concentration in Soybean. Crop Sci. 40: 1438.

Severin, A. J., S. B. Cannon, M. M. Graham, D. Grant, and R. C. Shoemaker, 2011 Changes in twelve homoeologous genomic regions in soybean following three rounds of polyploidy. Plant Cell 23: 3129–3136.

Shao, Z.-Q., Y.-M. Zhang, Y.-Y. Hang, J.-Y. Xue, G.-C. Zhou *et al.*, 2014 Long-Term Evolution of Nucleotide-Binding Site-Leucine-Rich Repeat (NBS-LRR) Genes: Understandings Gained From and Beyond the Legume Family. Plant Physiol. 166: 217–234.

Shen, X., Z.-Q. Liu, A. Mocoeur, Y. Xia, and H.-C. Jing, 2015 PAV markers in Sorghum bicolour: genome pattern, affected genes and pathways, and genetic linkage map construction. Theor. Appl. Genet. 128: 623–37.

Shoemaker, R. C., J. Schlueter, and J. J. Doyle, 2006 Paleopolyploidy and gene duplication in soybean and other legumes. Curr. Opin. Plant Biol. 9: 104–9.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–132.

Singer, T., and E. Burke, 2003 High-throughput TAIL-PCR as a tool to identify DNA flanking insertions. Methods Mol Biol 236: 241–272.

Song, Q., D. L. Hyten, G. Jia, C. V Quigley, E. W. Fickus *et al.*, 2013 Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8: e54985.

Song, Q., D. L. Hyten, G. Jia, C. V Quigley, E. W. Fickus *et al.*, 2015 Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 5: 1999–2006.

Srivastava, A., V. Philip, I. Greenstein, L. Rowe, M. Barter *et al.*, 2014 Discovery of transgene insertion sites by high throughput sequencing of mate pair libraries. BMC Genomics 15: 367.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley *et al.*, 2007 Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–53.

Stupar, R. M., and J. E. Specht, 2013 Insights from the soybean (Glycine max and Glycine soja) genome: past, present, and future (Donald L. Sparks, Ed.). Adv. Agron. 118: 177–204.

Sutton, T., U. Baumann, J. Hayes, N. C. Collins, B. J. Shi *et al.*, 2007 Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318: 1446–1449.

Svitashev, S. K., and D. A. Somers, 2002 Characterization of transgene loci in plants using FISH: A picture is worth a thousand words. Plant Cell. Tissue Organ Cult. 69: 205–214.

Swanson-Wagner, R. A., S. R. Eichten, S. Kumari, P. Tiffin, J. C. Stein *et al.*, 2010 Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20: 1689–1699.

Szpiech, Z. A., M. Jakobsson, and N. A. Rosenberg, 2008 ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics 24: 2498–504.

Takken, F., and M. Rep, 2010 The arms race between tomato and Fusarium oxysporum. Mol. Plant Pathol. 11: 309–14.

Tang, H., E. Lyons, B. Pedersen, J. C. Schnable, A. H. Paterson *et al.*, 2011 Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12: 102.

Tattini, L., R. D'Aurizio, and A. Magi, 2015 Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front. Bioeng. Biotechnol. 3: 92.

Tax, F. E., and D. M. Vernon, 2001 T-DNA-Associated Duplication/Translocations in Arabidopsis. Implications for Mutant Analysis and Functional Genomics. Plant Physiol. 126: 1527–1538.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini *et al.*, 2005 Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. U. S. A. 102: 13950–13955.

Thornton, K. R., A. J. Foran, and A. D. Long, 2013 Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. 9: e1003258.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14: 178–192.

Tian, D., M. B. Traw, J. Q. Chen, M. Kreitman, and J. Bergelson, 2003 Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. Nature 423: 74–7.

Tiffin, P., and J. Ross-Ibarra, 2014 Advances and limits of using population genetics to understand local adaptation. Trends Ecol. Evol. 29: 673–680.

Todd, J. J., and L. O. Vodkin, 1996 Duplications That Suppress and Deletions That Restore Expression from a Chalcone Synthase Multigene Family. Plant Cell 8: 687–699.

Turner, S. D., 2014 qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots: Cold Spring Harbor Labs Journals.

Tuteja, J. H., and L. O. Vodkin, 2008 Structural features of the endogenous CHS silencing and target loci in the soybean genome. Crop Sci. 48.:

Tuteja, J. H., G. Zabala, K. Varala, M. Hudson, and L. O. Vodkin, 2009 Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in glycine max seed coats. Plant Cell 21: 3063–3077.

Voytas, D. F., 2013 Plant genome engineering with sequence-specific nucleases. Annu. Rev. Plant Biol. 64: 327–50.

Wang, D., B. W. Diers, P. R. Arelli, and R. C. Shoemaker, 2001 Loci underlying resistance to Race 3 of soybean cyst nematode in Glycine soja plant introduction 468916. Theor. Appl. Genet. 103: 561–566.

Wang, J., C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley *et al.*, 2011 CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods 8: 652–4.

Wang, M., C. R. Beck, A. C. English, Q. Meng, C. Buhay *et al.*, 2015a PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC Genomics 16: 214.

Wang, Y., G. Xiong, J. Hu, L. Jiang, H. Yu *et al.*, 2015b Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nat. Genet. 47: 944–948.

Wang, Y., J. Lu, S. Chen, L. Shu, R. G. Palmer *et al.*, 2014 Exploration of presence/absence variation and corresponding polymorphic markers in soybean genome. J. Integr. Plant Biol. 56: 1009–1019.

Weber, N., C. Halpin, L. C. Hannah, J. M. Jez, J. Kough *et al.*, 2012 Crop Genome Plasticity and Its Relevance to Food and Feed Safety of Genetically Engineered Breeding Stacks. Plant Physiol. 160: 1842–1853.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-Statistics for the Analysis of Population Structure. Evolution (N. Y). 38: 1358–1370.

Wingen, L. U., T. Munster, W. Faigl, W. Deleu, H. Sommer *et al.*, 2012 Molecular genetic basis of pod corn (Tunicate maize). Proc. Natl. Acad. Sci. U. S. A. 109: 7115–7120.

Wright, S., 1949 The Genetical Structure Of Populations. Ann. Eugen. 15: 323–354.

Wu, X., C. Ren, T. Joshi, T. Vuong, D. Xu *et al.*, 2010 SNP discovery by high-throughput sequencing in soybean. BMC Genomics 11: 469.

Xie, Z., D. Li, L. Wang, F. D. Sack, and E. Grotewold, 2010 Role of the stomatal development regulators FLP/MYB88 in abiotic stress responses. Plant J. 64: 731–9.

Xu, K., X. Xu, T. Fukao, P. Canlas, R. Maghirang-Rodriguez *et al.*, 2006 Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442: 705–708.

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda *et al.*, 2005 A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17: 2859–72.

Yang, W.-Y., J. Novembre, E. Eskin, and E. Halperin, 2012 A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. 44: 725–731.

Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeau *et al.*, 2002 Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. Proc. Natl. Acad. Sci. U. S. A. 99: 6157–62.

Yoder, J. B., J. Stanton-Geddes, P. Zhou, R. Briskine, N. D. Young *et al.*, 2014 Genomic signature of adaptation to climate in Medicago truncatula. Genetics 196: 1263–1275.

Yu, P., C.-H. Wang, Q. Xu, Y. Feng, X.-P. Yuan *et al.*, 2013 Genome-wide copy number variations in Oryza sativa L. BMC Genomics 14: 649.

Zhang, C., S. Grosic, S. A. Whitham, and J. H. Hill, 2012 The requirement of multiple defense genes in soybean Rsv1-mediated extreme resistance to Soybean mosaic virus. Mol. Plant-Microbe Interact. 25: 1307–1313.

Zhang, D., Z. Wang, N. Wang, Y. Gao, Y. Liu *et al.*, 2014 Tissue culture-induced heritable genomic variation in rice, and their phenotypic implications. PLoS One 9: e96879.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42: 355–60.

Zhang, Z., L. Mao, H. Chen, F. Bu, G. Li *et al.*, 2015 Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. Plant Cell 27: 1595–604.

Zheng, L.-Y., X.-S. Guo, B. He, L.-J. Sun, Y. Peng *et al.*, 2011 Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol. 12: R114.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie *et al.*, 2012 A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326–8.

Zhou, H., B. Liu, D. P. Weeks, M. H. Spalding, and B. Yang, 2014 Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. Nucleic Acids Res. 42: 10903–14.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu *et al.*, 2015 Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33: 408–414.

Żmieńko, A., A. Samelak, P. Kozłowski, and M. Figlerowicz, 2014 Copy number polymorphism in plant genomes. Theor. Appl. Genet. 127: 1–18.

Appendix 1

Chapter 3 Supplemental:

SUPPLEMENTARY TABLES

Table S1. Results from CGH, breakpoint sequencing, TAIL-PCR, and resequencing of transgenic plants.

Transgenic Genotype	Construct	Data Types	Background	CGH-detected SV	T-DNA Location	T-DNA Direction	SV adjacent to T-DNA	No. Transgenes
WPT-384-1-1	TALEN	CGH, TAIL-PCR	Bert-MN-01	23,406 bp; Gm01 deletion	Gm07:35,729,576.. 35,729,766	+	NA	Likely 1
WPT-389-2-2	mPing Transposon	CGH, NGS, TAIL-PCR, Southern Blot	Bert-MN-01	125,228 bp; Gm11 deletion	Gm13:35,614,287.. 35,614,386	+	1,533 bp deletion + 37 bp deletion	1
WPT-301-3-13	GFP+RNAi Hairpin	CGH, TAIL-PCR, Southern Blot	Wm82-ISU-01	6,869 bp; Gm13 duplication	Gm04:2,694,961.. 2,694,962	-	NA	1
WPT-391-1-6	Magnesium Chelatase RNAi Hairpin	CGH, NGS	Bert-MN-01	7,854 bp; Gm19 deletion	Gm05:38,834,281.. 38,834,291	+	~1,200 bp deletion	1
WPT-312-5-126	Zinc Finger Nuclease	CGH, Southern Blot	Bert-MN-01	None	Untested	NA	NA	1

Table S2. Summary of data type, CGH design, and analysis method for Inter-cultivar Fast Neutron, and Transgenic genotypic classes.

	Inter-Cultivar	Fast Neutron	Transgenic
Original Experiment	Anderson *et al.*, 2014	Bolon *et al.*, 2014	Present Study
No. Genotypes Analyzed	41	35	5
Genotype Tested (Cy3)	SoyNAM Parent Accession	"No Phenotype" Mutant	Transformed Individual (T1)
Reference (Cy5)	Wm82-ISU-01	M92-220 - Long	Bert-MN-01 or Wm82-ISU-01
Data Types	CGH & Whole Genome Sequence	CGH	CGH
Analysis Method	Cross validation, visual analysis	Array based thresholds, visual analysis	Empirical thresholds, visual analysis
Experiment designed to detect	Genes affected by SV	SV induced genome-wide	SV induced genome-wide

Table S3. Summary of SV frequency in Inter-cultivar Fast Neutron, and Transgenic genotypic classes.

		Inter-Cultivar	Fast Neutron	Transgenic
Unique Up CNV Genes	Total genes in class	223	2118	2
	Maximum among genotypes	124	1568	2
	Median among genotypes	83	0	0
	Minimum among genotyps	45	0	0
Unique Down CNV Genes (homozygous or heterozygous deletions)	Total in class	1126	1231	4
	Maximum among genotypes	362	236	4
	Median among genotypes	244	12	0
	Minimum among genotyps	156	0	0
Up SV (homozgous duplications)	Total genic segments in class	117	9	1
	Mean Size	13,580 bp	2,447,335 bp	6,434 bp
	Median Size	3,182 bp	747,592 bp	6,434 bp
Down SV (homozygous or heterozygous deletion)	Total in class	547	49	1
	Mean Size	14,958 bp	515,051 bp	125,228 bp
	Median Size	2,775 bp	131,036 bp	125,228 bp

Table S4. Fast neutron genotypes from resequencing, all part of the forward screen family, Bolon *et al.*, 2014.

Soybase ID	Coded Name (This study)	Mean Coverage (BWA)	Putatitve deleted genes	No. chromos. with deleted genes	Putatitve duplicated genes	No. chromos. with duplicated genes	Soybase Family Name	CGH ID	M2 Family Name	Rad. Dose	Gen.	Mutant phenotype
M92-220.x1.04.WT	FN01 (M92-220 - Long)	64	-	-	-	-	-	-	-	-	-	-
FN0173217.03.09.01.M5	FN02	37	243	2	0	0	FN0173217	R32C17P18C09 #1 rep2	R32C17CSCW08YB	16 Gy	M5	High seed protein
FN0173217.03.09.01.M5	FN02	37	243	2	0	0	FN0173217	R32C17P18C09 #1 rep2	R32C17CSCW08YB	16 Gy	M5	High seed protein
FN0172932.09.08.01.M5	FN03	26	48	2	0	0	FN0172932	R29C32P13i08 #1 rep2	R29C32CSCW08YB	16 Gy	M5	High seed weight, low seed protein
FN0175143.05.06.01.M5	FN04	36	0	0	0	0	FN0175143	R51C43P26e06 #1 rep2	R51C43CSCW08YB	16 Gy	M5	High seed oil, high seed protein and oil
FN0171501.01.02.M4	FN05	31	56	2	2312	3	FN0171501	R15C01P33a02	R15C01DSCW08YB	32 Gy	M4	High seed protein
FN0131633.06.01.M4	FN06	34	7	2	0	0	FN0131633	3R16C33Cfr371aMN12	3R16C33CMN09NSFBV	16 Gy	M4	High seed oil, high seed protein and oil, high seed yield
FN0112228.06.02.	FN07	34	2	1	0	0	FN0112228	1R22C28Cfbr62aMN12	1R22C28CMN09	16 Gy	M5	High seed oil, low seed protein, high

01.M5									NSFBV			seed yield, late maturity, short, bushy, and indeterminate
FN0112885.02.06.03.M5	FN08	39	0	0	6	2	FN0112885	1R28C85Cbfr55cMN12	1R28C85CMN09NSFBV	16 Gy	M5	High seed oil, low seed protein, high seed yield, late maturity, short, bushy, and indeterminate
FN0163764.04.01.M4	FN09	22	92	3	934	2	FN0163764	6R37C64Ddr229aMN12 rep2	6R37C64DMN09NSFBV	32 Gy	M4	Stunted, short internodes, short petiole, slightly lanceolate leaves, early maturity, determinate
FN0164160.03.02.01.01.M6	FN10	32	290	3	1	1	FN0164160	6R41C60Dcbar163aMN12	6R41C60DMN09NSFBV	32 Gy	M6	Seed composition mutant, small plant, slightly chlorotic, slightly rugose, slightly tawny pubescence
FN0175501.x2.02.01.M5	FN11	30	6	2	0	0	FN0175501	GMGC2ba	R55C01CSCW08YB	16 Gy	M5	Short trichomes

Table S5. Summary of SNPs and frequency in subsample of Fast Neutron and Transgenic experiments.

	FN01 M92- 220	FN02	FN03	FN04	FN06	FN07	FN08	FN11	FN05	FN09	FN10	Bert -1	Bert -2	WPT389 -2-2	WPT391 -1-6
Dosage	NA	16 Gy	16 Gy	16 Gy	16 Gy	16 Gy	16 Gy	16 Gy	32 Gy	32 Gy	32 Gy	-	-	-	-
Generation	-	M5	M5	M5	M4	M5	M5	M5	M4	M4	M6	-	-	T1	T1
Homozygous Substitutions	41	45	42	41	49	58	62	44	76	50	73	2	1	18	2
Genic:															
Coding	2	1	4	2	1	5	16	4	3	1	6	0	0	0	0
Non-Coding	3	5	2	5	9	21	10	4	7	1	4	0	1	1	0
Non-Genic	36	39	36	34	39	32	36	36	66	48	63	2	0	17	2
Ti:Tv Ratio	-	2.4	1.7	1.9	1.2	3.0	0.8	1.2	1.3	1.9	1.4	-	-	1.5	0

Table S6. Genotypes and Hemizigous regions for developing empirical thresholds.

Genotype	Segment Type	Chromosome	Average Log2Ratio	No. probes	Region Start	Region Stop	Region Size	Used as Universal Threshold
3R16C33Cfr371aMN12	Hemizygous Deletion	GM16	-0.525706452	899	8161171	8737551	576380	Yes
5R15C49Dcdr81aMN12	Deletion	GM07	-0.589188374	2116	28900343	30975759	2075416	
6R41C60Dcbar163aMN12	Deletion	GM04	-0.731	2640	42480798	43845671	1364874	
R02C28-7-35-1	Deletion	GM07	-0.657297156	5662	24452904	31229508	6776604	
R07C12-6-41-1	Deletion	GM15	-0.7267	3332	43545233	46011969	2466737	
R32C17P18C09 #1	Deletion	GM06	-0.634165475	5593	22989683	29272095	6282412	
R32C17P18C09 #1	Deletion	GM06	-0.584217327	9067	31864518	39994964	8130446	
6R41C60Dcbar163aMN12	Duplication	GM04	0.4252	2955	43846110	45384387	1538278	
R02C28-7-35-1	Duplication	GM15	0.376373176	74406	1	50938913	50938912	
R07C12-6-41-1	Duplication	GM15	0.384310595	60774	1	42984752	42984751	
R07C12-6-41-1	Duplication	GM15	0.42241886	8033	46567523	50938913	4371390	
R07C12-6-41-1	Duplication	GM16	0.8113	3511	35311624	37131684	1820061	
R15C01P33a02	Duplication	GM04	0.390922236	28090	81	16866782	16866701	
R15C01P33a02	Duplication	GM04	0.3484	3272	19534674	23730015	4195342	Yes
R15C01P33a02	Duplication	GM04	0.3569	1885	29586547	31624484	2037938	
R15C01P33a02	Duplication	GM08	0.356043871	8256	29455508	36860787	7405279	
R15C01P33a02	Duplication	GM08	0.396	17935	36862182	46994705	10132524	
R15C01P33a02	Duplication	GM18	0.3598	2284	24740911	27269292	2528382	
RP69dm4MNS12	Duplication	GM03	0.507063898	3166	28675179	30920725	2245546	

Table S7. Genotypes examined by CGH.

Class	Genotype Tested	Hybridized to Genotype:	Publication	Radiation Dose	Generation	GEO Series	GEO Accession
Transgenic	WPT0389-2-2_mPingline	Bert-MN-01	This study	-	T1	GSE73596	GSM1898745
Transgenic	WPT0391-1-6_MinnGold_hp	Bert-MN-01	This study	-	T1	GSE73596	GSM1898746
Transgenic	WPT0384-1-1_TALEN_Dcl2b	Bert-MN-01	This study	-	T1	GSE73596	GSM1898744
Transgenic	WPT_312_5_126_ZFN	Bert-MN-01	This study	-	T1	GSE73596	GSM1898743
Transgenic	WPT_301_3_13_GFP_RNAi Hairpin	Wm82-ISU-01	This study	-	T1	GSE73596	GSM1898742
Control	Bert-MN-01	Bert-MN-01	This study	-	-	GSE73596	GSM1898747
Control	Williams	Wm82-ISU-01	This study	-	-	GSE73596	GSM1898748
NAM parent	TN05-3027 _NAM 02	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359718
NAM parent	4J105-3-4_NAM 03	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359719
NAM parent	5M20-2-5-2_NAM 04	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359720
NAM parent	CL0J095-4-6_NAM 05	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359721
NAM parent	CL0J173-6-8_NAM 06	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359722
NAM parent	HS6-3976 _NAM 08	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359723
NAM parent	Prohio_NAM 09	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359724
NAM parent	LD00-3309_NAM 10	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359725
NAM parent	LD01-5907 _NAM 11	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359726

NAM parent	LD02-4485_NAM 12	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359727
NAM parent	LD02-9050 _NAM 13	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359728
NAM parent	Magellan_NAM 14	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359729
NAM parent	Maverick_NAM 15	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359730
NAM parent	S06-13640 _NAM 17	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359731
NAM parent	NE3001_NAM 18	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359732
NAM parent	Skylla _NAM 22	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359733
NAM parent	U03-100612 _NAM 23	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359734
NAM parent	LG03-2979 _NAM 24	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359735
NAM parent	LG03-3191 _NAM 25	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359736
NAM parent	LG04-4717 _NAM 26	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359737
NAM parent	LG05-4292_NAM 27	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359738
NAM parent	LG05-4317_NAM 28	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359739
NAM parent	LG05-4464_NAM 29	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359740
NAM parent	LG05-4832_NAM 30	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359741
NAM parent	LG90-2550 _NAM 31	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359742
NAM parent	LG92-1255_NAM 32	Wm82-ISU-01	Anderson *et al.*,	-	-	GSE56351	GSM1359743

			2014				
NAM parent	LG94-1128 _NAM 33	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359744
NAM parent	LG94-1906_NAM 34	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359745
NAM parent	LG97-7012_NAM 36	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359746
NAM parent	LG98-1605 _NAM 37	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359747
NAM parent	LG00-3372 _NAM 38	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359748
NAM parent	LG04-6000 _NAM 39	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359749
NAM parent	PI 398.881_NAM 40	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359750
NAM parent	PI 427.136_NAM 41	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359751
NAM parent	PI 437.169B_NAM 42	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359752
NAM parent	PI 507.681B_NAM 46	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359753
NAM parent	PI 518.751_NAM 48	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359754
NAM parent	PI 561.370_NAM 50	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359755
NAM parent	PI 404.188A _NAM 54	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359756
NAM parent	PI 574.486_NAM 64	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359757
NAM parent	IA3023_NAM Universal Parent	Wm82-ISU-01	Anderson *et al.*, 2014	-	-	GSE56351	GSM1359758
Fast Neutron	1R03C38Cbr290aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402584

Fast Neutron	1R04C95Cbr291cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402585
Fast Neutron	1R12C21Ccr292cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402586
Fast Neutron	1R19C96Cfr293aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402587
Fast Neutron	1R23C51Cdr294aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402589
Fast Neutron	1R25C46Car295bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402590
Fast Neutron	1R28C71Cdr296cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402591
Fast Neutron	1R36C46Ccr297bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402593
Fast Neutron	2R01C05Ccr298aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402641
Fast Neutron	2R01C66Cfr299cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402642
Fast Neutron	2R02C47Cbr300aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402643
Fast Neutron	2R06C87Ccr301bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402644
Fast Neutron	2R07C12Cjr302aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402645
Fast Neutron	2R10C37Cdr303bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402646
Fast Neutron	2R11C31Cjr304cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402647
Fast Neutron	2R11C55Cdr305bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402648
Fast Neutron	2R25C69Ccr306cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402649
Fast Neutron	2R39C51Car307aMN11	M92-220 -	Bolon *et al.*,	16 Gy	M4	GSE58172	GSM1402655

		Long	2014				
Fast Neutron	2R43C67Cbr308bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402656
Fast Neutron	2R47C02Ccr309cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402658
Fast Neutron	2R47C48Car310aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402659
Fast Neutron	3R03C42Clr311bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402660
Fast Neutron	3R11C39Cbr312cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402661
Fast Neutron	3R23C38Car313aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402663
Fast Neutron	3R23C70Cgr314bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402664
Fast Neutron	3R27C90Cer315cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402665
Fast Neutron	3R33C61Ccr316aMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402666
Fast Neutron	4R02C19Car317bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402670
Fast Neutron	4R03C13Cbr318cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402671
Fast Neutron	4R09C72Dar319cMN11	M92-220 - Long	Bolon *et al.*, 2014	32 Gy	M4	GSE58172	GSM1402674
Fast Neutron	4R17C66Dbr320bMN11	M92-220 - Long	Bolon *et al.*, 2014	32 Gy	M4	GSE58172	GSM1402676
Fast Neutron	4R38C67Cbr321cMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402678
Fast Neutron	5R16C69Abr322aMN11	M92-220 - Long	Bolon *et al.*, 2014	4 Gy	M4	GSE58172	GSM1402688
Fast Neutron	5R28C09Cdr323aMn11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402689

Fast Neutron	5R29C21Chr324bMN11	M92-220 - Long	Bolon *et al.*, 2014	16 Gy	M4	GSE58172	GSM1402690

Table S8. Genotyping Primer Sequences.

SV location and Background	Primer Name	Sequence	Backup Primer
Chromosome 1, WPT_384-1-1	F_Deletion	AGTAGCGGAACTGGTGTGGT	TTTGTCATCCTCGTCGTTTG
Chromosome 1, WPT_384-1-1	F_WildType	GTTTGTTGTGGAGTGTTAGC	
Chromosome 1, WPT_384-1-1	Reverse	CACAAAGGCCACAAATTGAA	CATGCACAACGTGGTCTTTC
Chromosome 11, WPT_389-2-2	F_Deletion	CACAAACTTGGACTGCTGGA	
Chromosome 11, WPT_389-2-2	F_WildType	GGAGTGCAGGTTGCTTGAGC	
Chromosome 11, WPT_389-2-2	Reverse	TAGTTTTCGTCGGCAAAAGG	
Chromosome 13, WPT_301-3-13	F_Duplication	GCTCAATTTGGTCCTTTCCA	
Chromosome 13, WPT_301-3-13	F_WildType	GCATGAAAGGGTATAGGAAGG	
Chromosome 13, WPT_301-3-13	Reverse	GTCTAGAACCCTATCCGTGCAC	
Chromosome 19, WPT_391-1-6	F_Deletion	GTGTAGTAAGAAAATGCTCACC	
Chromosome 19, WPT_391-1-6	Reverse	GCCATCAATGCCTCAGAAAC	

SUPPLEMENTARY FIGURES

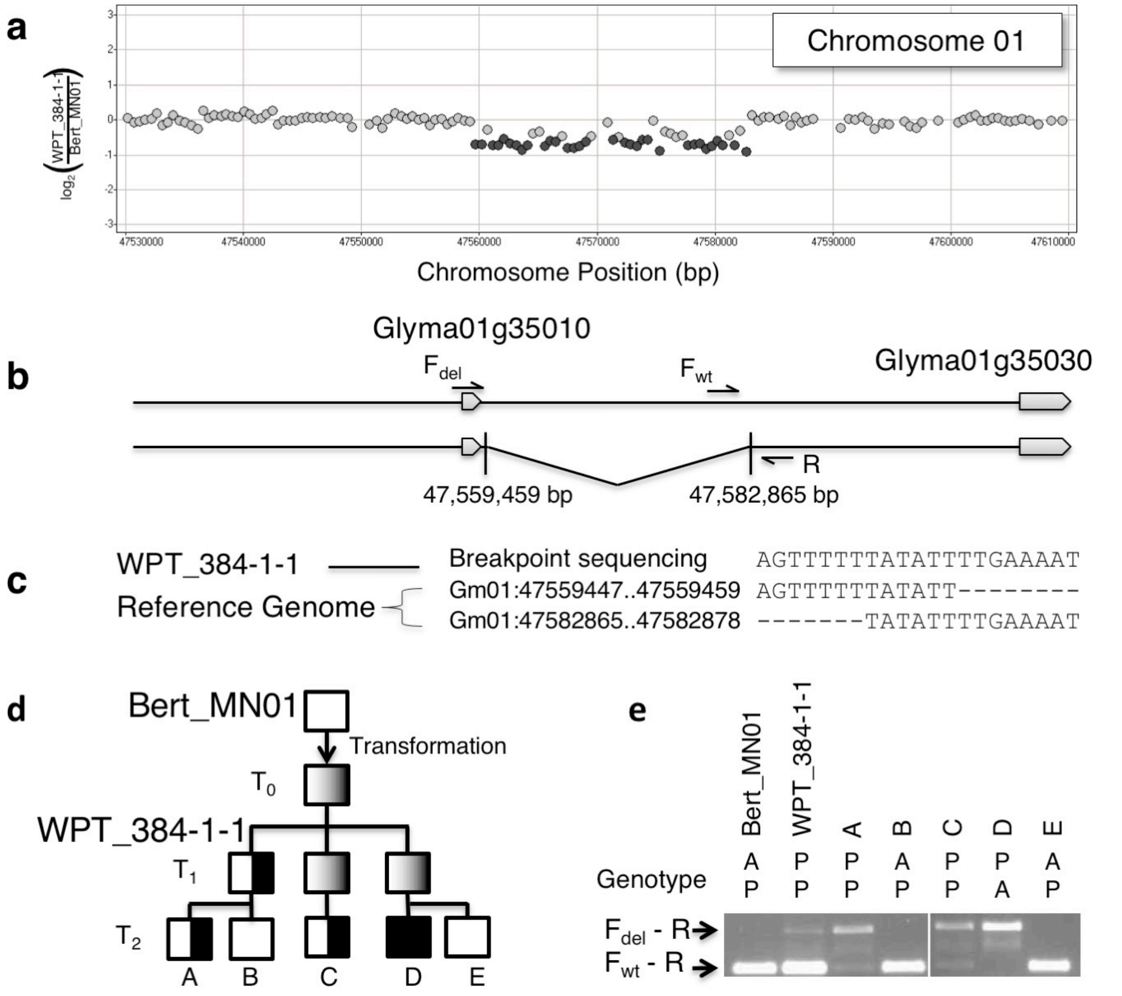


Figure S1. A novel deletion detected on chromosome 01 in transgenic line WPT_384-1-1. **(a)** Plot of CGH data for the transgenic line versus Bert-MN01, zoomed in on the chromosome 01 deletion. Probes are plotted as dots corresponding to the $\log_2$ ratio from the CGH array. Dark gray dots represent probes within segments that exceed the empirical threshold. The amplitude of the trough indicates a putative hemizygous deletion, wherein one homologous chromosome harbors the deletion while the other homolog is normal (wild-type). **(b)** Graphical interpretation of the deletion found in WPT_384-1-1, showing one normal and one deletion-bearing chromosome. Black arrows indicate orientation and location of genotyping primers F_{del}, F_{wt}, and R. **(c)** Sequence data from the breakpoint junction shows six base pairs of homology on either end of the breakpoint. Pedigree **(d)** and genotyping data **(e)** show the deletion's stability across generations. Electrophoresis gels demonstrate genotyping the deletion band (642 bp) and wild type band (252 bp) for individuals labeled in the pedigree. Bands scored as present (P) or absent (A).

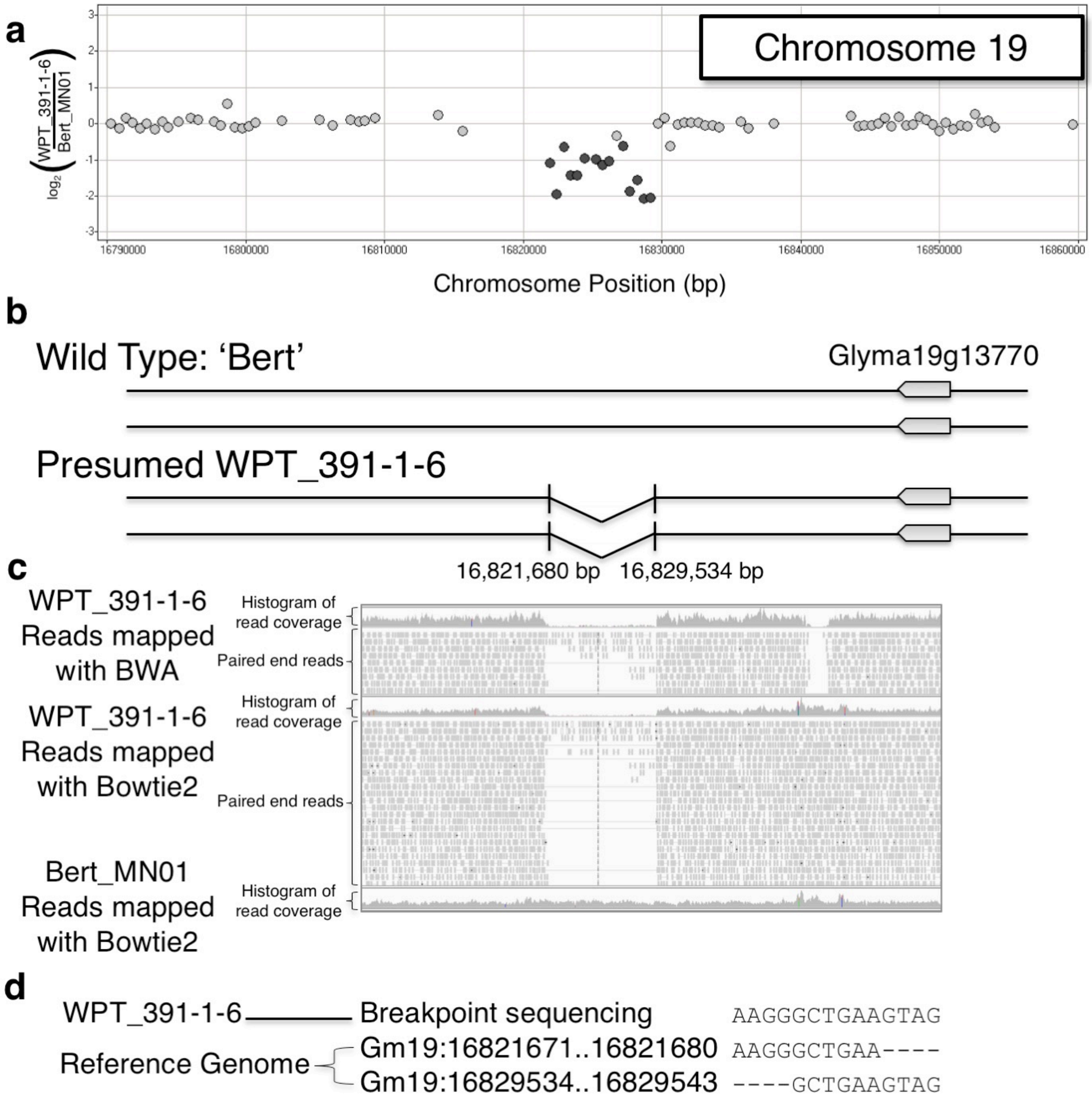


Figure S2. Deletion on chromosome 19 in transgenic line WPT_391-1-6. **(a)** Plot of CGH data for the transgenic line versus 'Bert_MN01', zoomed in on the chromosome 19 deletion. Probes are plotted as dots corresponding to the $\log_2$ ratio from the CGH array. Dark gray dots represent probes within significant segments that exceed the empirical threshold. **(b)** Graphical interpretation of the reference wild type and the homozygous deletion found in WPT_391-1-6. **(c)** Read coverage through multiple mapping methods in this region. Low read coverage confirms the CGH detected deletion. Multiple paired ends reads spanning the breakpoint when mapped with Bowtie2 also confirm the deletion. **(d)** Breakpoint sequencing suggesting microhomology based repair. Deletion size is 7,854 bp in size.

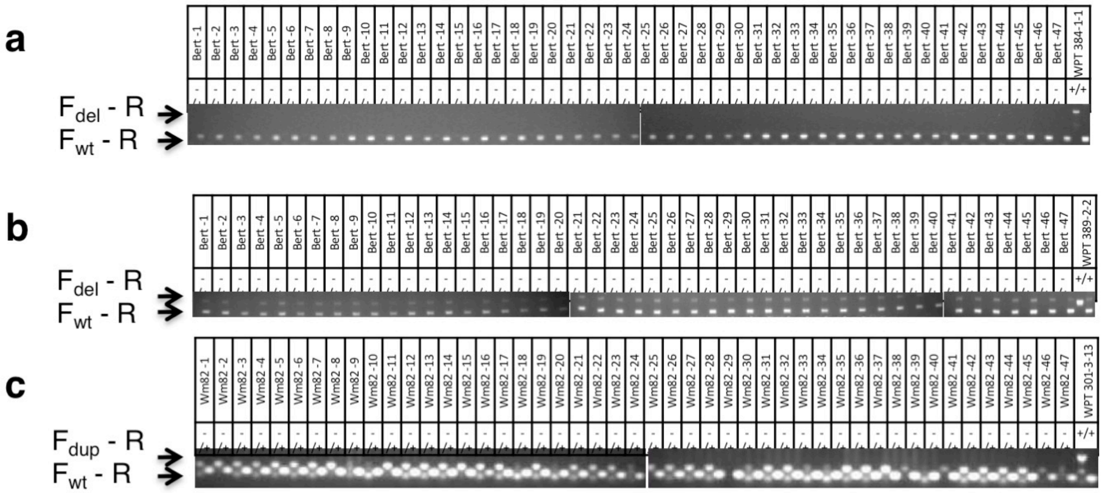


Figure S3. Genotyping 47 individuals from the transformation varieties' (Bert or Williams 82) GRIN database collection to test for intracultivar variation. Positive control is in the last column. **(a)** The deletion on chromosome 1 found in WPT_384-1-1, a transcription activator-like effector nuclease construct in a 'Bert_MN_01' background, does not show up in any of the Bert individuals sampled. **(b)** The deletion on chromosome 11 found in WPT_398-2-2, an mPing transposon construct in a 'Bert_MN_01' background, does not show up in any of the Bert individuals sampled. **(c)** The duplication on chromosome 13 found in WPT_301-3-13, a GFP and RNAi hairpin construct in 'Wm82_ISU_01' background, does not show up in any of the Williams82 individuals sampled. None of these three SV can be attributed to intracultivar variation.

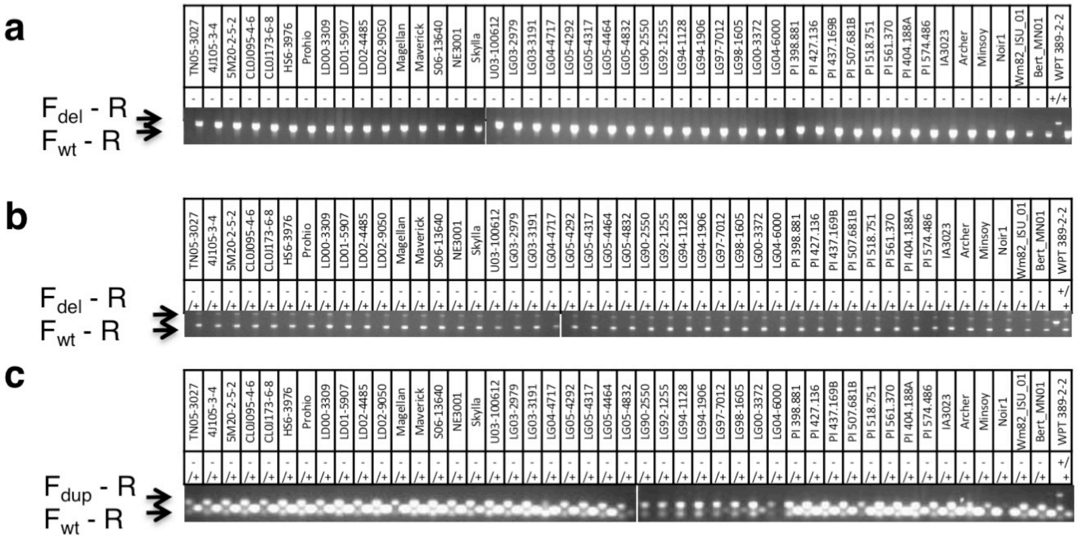


Figure S4. Genotyping diverse lines including the 41 SoyNAM parents, cultivars Archer, Minsoy, and Noir1, sublines 'Bert_MN_01', and 'Wm82_ISU_01,' for previous evidence of transformation induced SV. Positive control is in the last column. **(a)** The deletion on chromosome 1 found in WPT_384-1-1, a transcription activator-like effector nuclease construct in a 'Bert-MN01' background, does not show up in any of the diverse individuals sampled. **(b)** The deletion on chromosome 11 found in WPT_398-2-2, an mPing transposon construct in a 'Bert-MN01' background , does not show up in any of the diverse individuals sampled. **(c)** The duplication on chromosome 13 found in WPT_301-3-13, a GFP and RNAi hairpin construct in 'Wm82-ISU01' background, does not show up in any of the diverse individuals sampled. None of these SV were a result of contamination.

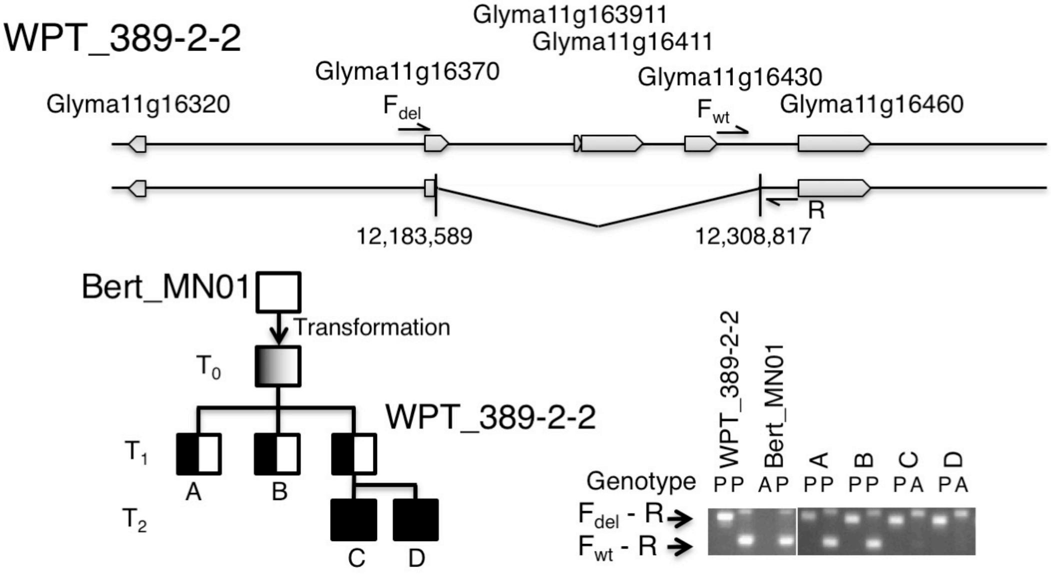


Figure S5. Novel deletion on chromosome 11 in transgenic line WPT_389-2-2. Black arrows indicate orientation and location of genotyping primers F$_{del}$, F$_{wt}$, and R. Pedigree and genotyping data show the deletion's stability across generations. Electrophoresis gels demonstrate genotyping the deletion band (406 bp) and internal band (285 bp) for individuals labeled in the pedigree. Bands scored as present (P) or absent (A).

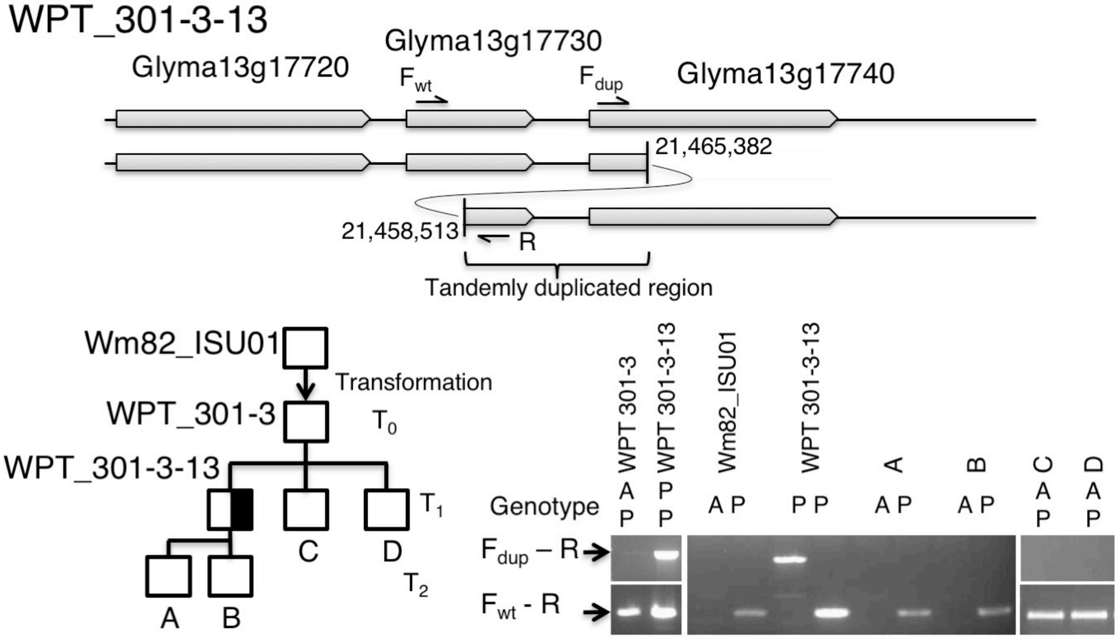


Figure S6. Novel duplication on chromosome 13 in transgenic line WPT_301-3-13. Black arrows indicate orientation and location of genotyping primers F_{dup}, F_{wt}, and R. Pedigree and genotyping data show the duplication's lack of segregation across generations. In this case the F_{wt}-R primer combo can only confirm DNA quality and can not aid in genotyping heterozygotes. Electrophoresis gels show the duplication band (929 bp) and internal band (329 bp) for individuals labeled in the pedigree. Bands scored as present (P) or absent (A). This duplication is real but appears to have been induced naturally and not by the transformation process.

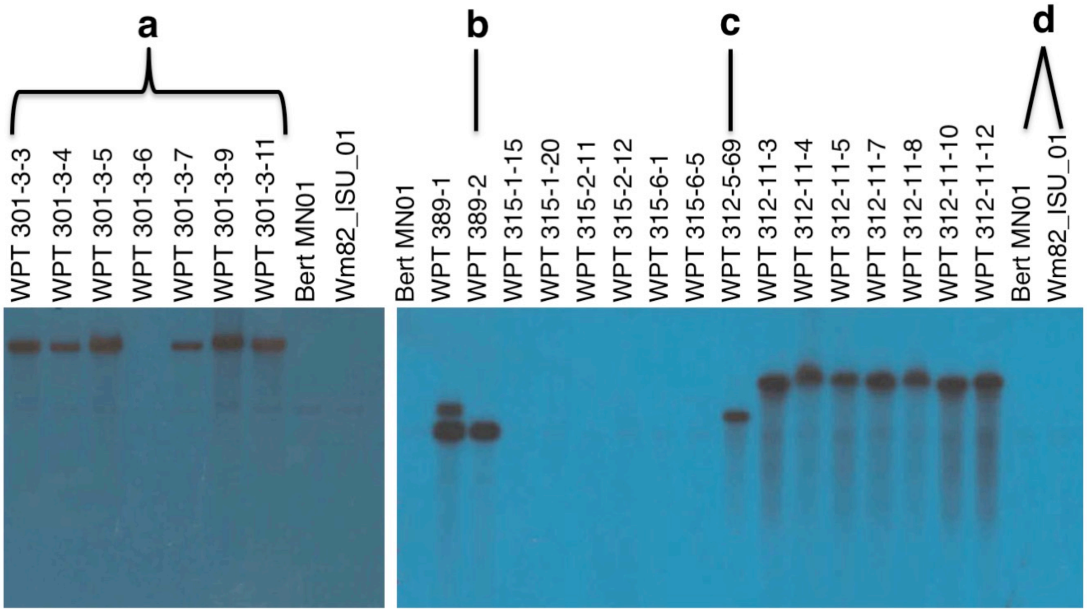


Figure S7. Southern blot analysis of *Hind*III digested genomic DNA. Probe located in the BAR gene in each vector background. **(a)** Siblings to WPT_301-3-12 segregate for a single T-DNA insertion. **(b)** Parent of WPT 389-2-2 has a single T-DNA insertion. **(c)** Sibling of WPT 312-5-126 has a single T-DNA insertion. **(d)** Controls show no T-DNA in Bert MN01 and Wm82_ISU_01.

```
WPT_301-3-13, T-DNA Construct: GFP+RNAi Hairpin, Non-genic insertion
Breakpoint Sequencing        CACAATATATGAATGA
Left Border Sequence         CACAATATAT------
Gm04:2,695,263..2,695,270    --------ATGAATGA

WPT_391-1-6, T-DNA Construct: Magnesium Chelatase RNAi Hairpin, Non-genic insertion
Breakpoint sequencing        CCACAATATGTGTAAAG
Left Border Sequence         CCACAATATAT------
Gm05:38,834,281..38,834,291  ------TATGTGTAAAG
                                                                  12 bp deleted
WPT_384-1-1, T-DNA Construct: TALEN, Non-genic insertion
Breakpoint Sequencing        GCCCGTCTCAATTTGTGAGCCAATCACGCTAGAAGGT------------TGAGTTATA
Left Border Sequence         GCCCGTCTC-ACTGGTGA----------------------------------------
Gm07:35,729,562..35,729,615  ----GTCCAAATTTGTGAGCCAATCACGCTAGAAGGTCACACATGCTTCTCTGCTATA
```

Figure S8. Microhomology discovered through breakpoint sequencing integration site for three of the transgenic lines. Homology occurs between left border and genomic DNA. Nucleotide sites in bold are not homologous.

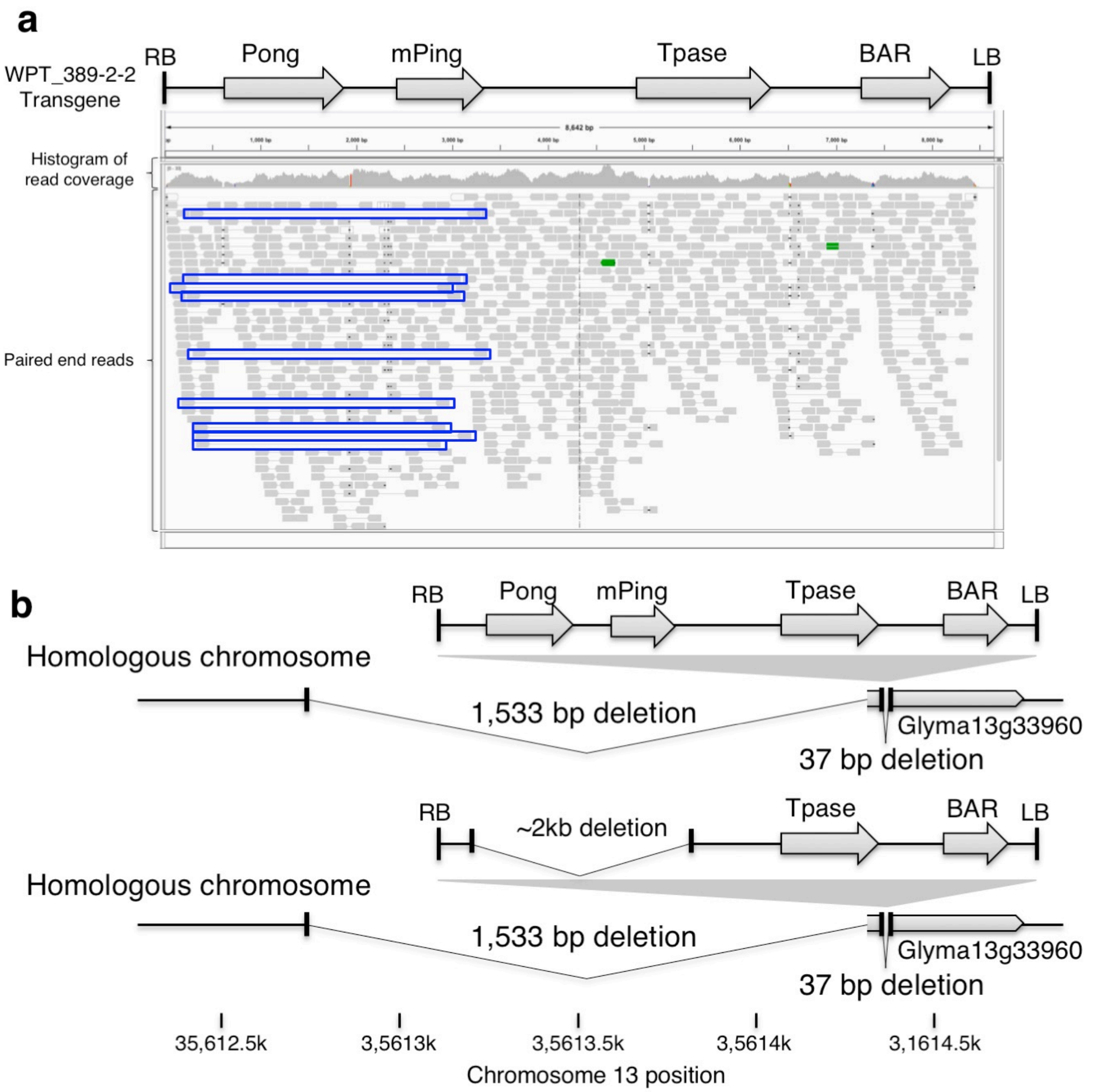


Figure S9. Transgene on chromosome 13 in transgenic line WPT_389-2-2. **(a)** WPT_389-2-2 construct contained four primary elements between the left and right Borders: Pong, mPing, Tpase, and BAR. Visual display in IGV of paired end reads mapped to the transgene using Bowtie. One of the transgene copies has lost approximately 2kb corresponding to the Pong and mPing regions of the T-DNA. The nine paired end reads suggesting this deletion are outlined in blue boxes. **(b)** Graphical interpretation of the homologous chromosomes at the T-DNA insertion site. The top chromosome represents the full T-DNA version while the other has a copy with a vacated mPing-Pong component. There was no evidence of this mPing-Pong fragment reinserting in the genome.

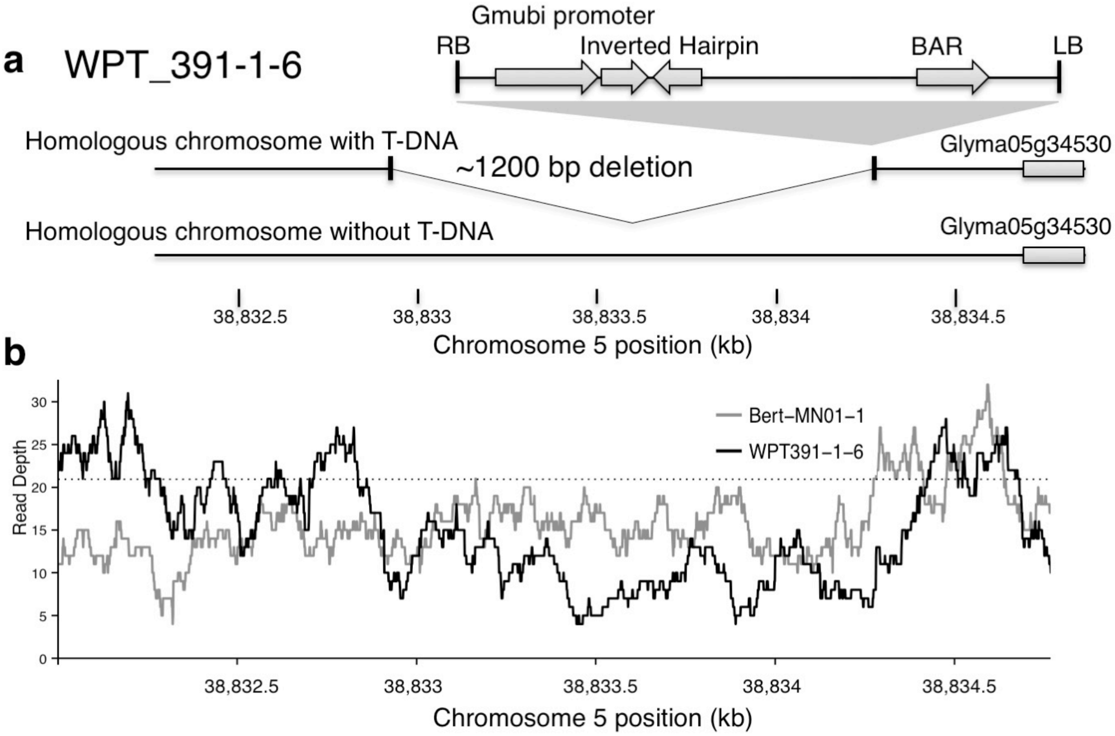


Figure S10. Transgene insertion heterozygous on chromosome 05 in transgenic line WPT_391-1-6. **(a)** Proposed location and orientation of transgene and deletion in genomic DNA in WPT_391-1-6. Transgene and deletion are on the same chromosome and are heterozygous. Construct used for WPT_391-1-6 contained three primary elements between the left and right borders: Gmubi promoter, inverted hairpin, and BAR. **(b)** Read depth coverage in this region when mapping with BWA suggesting a reduction in coverage in WPT_391-1-6 associated with a hemizygous deletion. The average coverage was 21x in both Bert and WPT_391-1-6 BWA alignments. This T-DNA insertion appears to have induced a 1.2 kb deletion in the process of integration.

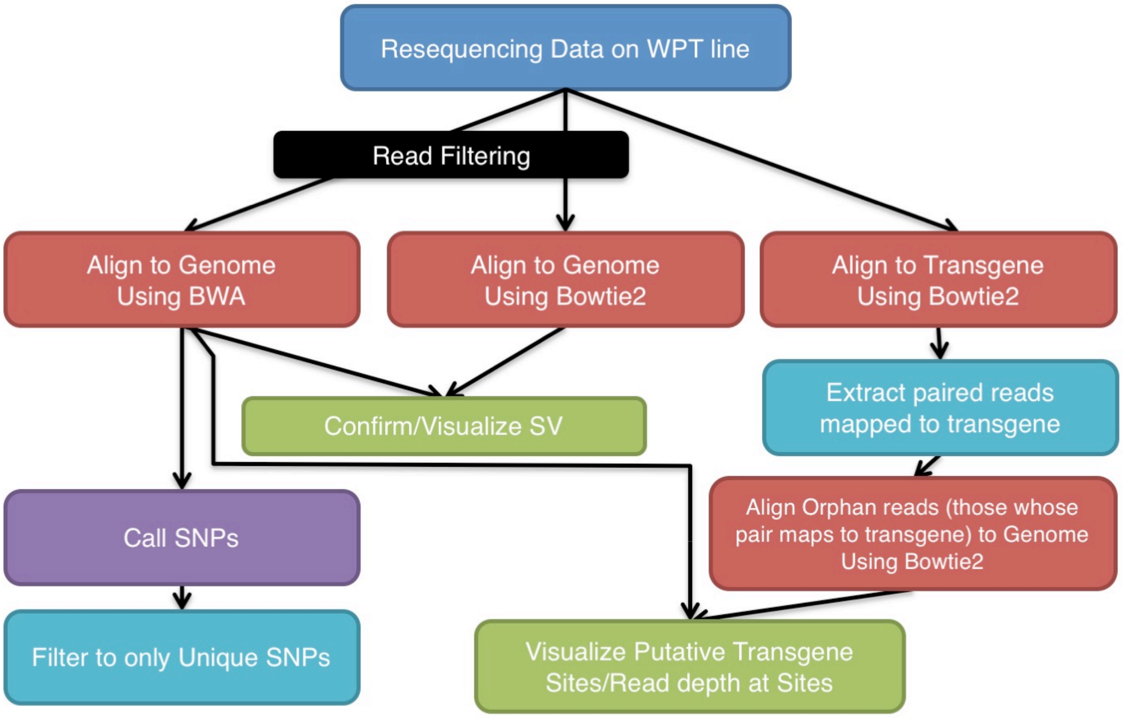

Resequencing Data on WPT line
Read Filtering
Align to Genome Using BWA
Align to Genome Using Bowtie2
Align to Transgene Using Bowtie2
Confirm/Visualize SV
Extract paired reads mapped to transgene
Align Orphan reads (those whose pair maps to transgene) to Genome Using Bowtie2
Call SNPs
Filter to only Unique SNPs
Visualize Putative Transgene Sites/Read depth at Sites

Figure S11. Pipeline for utilizing resequencing data in this study. This data contributed to confirming CGH discovered SV, calling SNPs, and localizing transgenes.

Appendix 2

Chapter 4 Supplemental:

SUPPLEMENTARY TABLES

Table S1: PI Number, latitude, longitude, country, and STRUCTURE-identified population of origin for each of the accessions used in this study.

PI	Country_of_origin	latitude	longitude	Structure Group	Color
PI339731	KOR	37.91	128.04	1	Green2
PI339732	KOR	37.31	128.535	1	Green2
PI339733	KOR	38.15	127.3	1	Green2
PI349647	KOR	37.2840004	127.0189972	1	Green2
PI366119	JPN	34.8500023	136.9166718	2	Red
PI366120	JPN	39.5333328	140.3833313	2	Red
PI366121	JPN	37.459611	139.841056	2	Red
PI366122	JPN	37.459611	139.841056	2	Red
PI366124	JPN	34.233333	133.783333	2	Red
PI366125	JPN	36.0314706	139.5339203	2	Red
PI378683	JPN	36.5333328	136.6166687	2	Red
PI378685	JPN	34.0333328	132.8500061	2	Red
PI378688	JPN	34.583334	135.6166687	2	Red
PI378689	JPN	37.1000023	138.2499924	2	Red
PI378690	JPN	33.2000008	130.3666687	2	Red
PI378691	JPN	31.458333	131.2333374	2	Red
PI378692	JPN	39.7	141.2	2	Red
PI378698	JPN	35.4500008	138.8500061	2	Red
PI378702	JPN	39.7	141.2	2	Red
PI407018	JPN	39.5437	140.298	2	Red
PI407019	JPN	39.5437	140.298	2	Red
PI407020	JPN	39.549	140.36	2	Red
PI407021	JPN	39.549	140.36	2	Red
PI407022	JPN	39.565	140.4025	2	Red
PI407023	JPN	39.565	140.4025	2	Red
PI407024	JPN	39.565	140.4025	2	Red
PI407025	JPN	39.5736	140.416	2	Red
PI407026	JPN	39.5736	140.416	2	Red
PI407027	JPN	39.5736	140.416	2	Red
PI407028	JPN	39.7230323	140.0671005	2	Red
PI407029	JPN	39.7230323	140.0671005	2	Red
PI407030	JPN	39.549	140.36	2	Red
PI407031	JPN	39.549	140.36	2	Red
PI407033	JPN	39.549	140.36	2	Red
PI407034	JPN	39.5	140.36	2	Red
PI407035	JPN	39.5	140.36	2	Red

PI407036	JPN	39.5	140.36	2	Red
PI407037	JPN	39.700069	140.730588	2	Red
PI407038	JPN	39.700069	140.730588	2	Red
PI407039	JPN	39.700069	140.730588	2	Red
PI407040	JPN	39.700069	140.730588	2	Red
PI407042	JPN	39.5736	140.416	2	Red
PI407043	JPN	39.5736	140.416	2	Red
PI407044	JPN	39.5736	140.416	2	Red
PI407045	JPN	39.7230323	140.0671005	2	Red
PI407046	JPN	39.7230323	140.0671005	2	Red
PI407047	JPN	39.7230323	140.0671005	2	Red
PI407048	JPN	39.7166285	141.1383963	2	Red
PI407049	JPN	39.7166285	141.1383963	2	Red
PI407050	JPN	39.7166285	141.1383963	2	Red
PI407052	JPN	39.7166285	141.1383963	2	Red
PI407053	JPN	36.1000023	137.9666672	2	Red
PI407055	JPN	35.0030033	138.0047607	2	Red
PI407056	JPN	34.8500023	136.9166718	2	Red
PI407057	JPN	34.8500023	136.9166718	2	Red
PI407058	JPN	34.8500023	136.9166718	2	Red
PI407059	JPN	34.8500023	136.9166718	2	Red
PI407060	JPN	34.8165116	136.9166718	2	Red
PI407061	JPN	34.8500023	136.9166718	2	Red
PI407068	JPN	34.7666683	136.8999939	2	Red
PI407069	JPN	34.7666683	136.8999939	2	Red
PI407070	JPN	34.7666683	136.8999939	2	Red
PI407071	JPN	34.9557	137.6066	2	Red
PI407072	JPN	34.9557	137.6066	2	Red
PI407073	JPN	34.9557	137.6066	2	Red
PI407074	JPN	34.9557	137.6066	2	Red
PI407076	JPN	34.9354819	137.2357177	2	Red
PI407077	JPN	34.9354819	137.2357177	2	Red
PI407080	JPN	34.7999992	136.8666687	2	Red
PI407081	JPN	34.7999992	136.8666687	2	Red
PI407082	JPN	34.7999992	136.8666687	2	Red
PI407083	JPN	34.9354819	137.2357177	2	Red
PI407085	JPN	34.9354819	137.2357177	2	Red
PI407087	JPN	34.7999992	134.9833374	2	Red
PI407088	JPN	34.7999992	134.9833374	2	Red
PI407089	JPN	34.7999992	134.9833374	2	Red
PI407090	JPN	34.7999992	134.9833374	2	Red
PI407091	JPN	34.8200043	135.1167297	2	Red
PI407092	JPN	35.1333332	134.9574211	2	Red
PI407094	JPN	34.8833332	135.1926954	2	Red
PI407095	JPN	34.9168238	135.2333374	2	Red
PI407096	JPN	34.9168238	135.2333374	2	Red

PI407097	JPN	34.757392	135.1554179	2	Red
PI407099	JPN	34.757392	135.1554179	2	Red
PI407100	JPN	34.757392	135.1554179	2	Red
PI407103	JPN	34.757392	135.1554179	2	Red
PI407104	JPN	34.757392	135.1554179	2	Red
PI407105	JPN	34.757392	135.1554179	2	Red
PI407107	JPN	34.757392	135.1554179	2	Red
PI407108	JPN	34.757392	135.1554179	2	Red
PI407109	JPN	34.757392	135.1554179	2	Red
PI407110	JPN	34.757392	135.1554179	2	Red
PI407111	JPN	35	135	2	Red
PI407112	JPN	35	135	2	Red
PI407113	JPN	35	135	2	Red
PI407114	JPN	35	135	2	Red
PI407115	JPN	34.9999962	134.9999924	2	Red
PI407116	JPN	35	135	2	Red
PI407117	JPN	35	135	2	Red
PI407118	JPN	35	135	2	Red
PI407121	JPN	35	135	2	Red
PI407124	JPN	35	135	2	Red
PI407125	JPN	34.9999962	134.9999924	2	Red
PI407126	JPN	34.6499977	133.9166718	2	Red
PI407128	JPN	32.816667	130.9	2	Red
PI407129	JPN	32.816667	130.9	2	Red
PI407130	JPN	32.816667	130.9	2	Red
PI407132	JPN	32.9999962	130.9999924	2	Red
PI407133	JPN	32.8166676	130.6999969	2	Red
PI407134	JPN	32.9132429	130.5862426	2	Red
PI407136	JPN	32.9132429	130.5862426	2	Red
PI407138	JPN	32.9132429	130.5862426	2	Red
PI407140	JPN	32.8907007	130.5862426	2	Red
PI407142	JPN	32.8907007	130.5862426	2	Red
PI407143	JPN	32.8907007	130.5862426	2	Red
PI407144	JPN	32.8907007	130.5862426	2	Red
PI407145	JPN	32.8907007	130.5862426	2	Red
PI407146	JPN	32.8907007	130.5862426	2	Red
PI407147	JPN	32.9222598	130.5862426	2	Red
PI407148	JPN	32.8336053	130.7166672	2	Red
PI407149	JPN	32.8336053	130.7166672	2	Red
PI407150	JPN	32.8336053	130.7166672	2	Red
PI407151	JPN	32.8336053	130.7166672	2	Red
PI407152	JPN	32.8336053	130.7166672	2	Red
PI407153	JPN	32.8336053	130.7166672	2	Red
PI407154	JPN	32.7999992	130.7166672	2	Red
PI407155	JPN	32.8792	130.743	2	Red
PI407156	JPN	35.4500008	138.8500061	2	Red

PI407157	JPN	35.6000023	140.1166687	2	Red
PI407159	KOR	37.2840004	127.1092097	1	Green2
PI407160	KOR	37.2840004	127.1092097	1	Green2
PI407161	KOR	37.2840004	127.1092097	1	Green2
PI407162	KOR	37.2840004	127.1092097	2	Red
PI407163	KOR	37.2840004	127.1092097	1	Green2
PI407164	KOR	37.2840004	127.1092097	1	Green2
PI407165	KOR	37.2840004	127.1092097	1	Green2
PI407166	KOR	37.2840004	127.1204862	1	Green2
PI407167	KOR	37.2840004	127.1204862	1	Green2
PI407168	KOR	37.2840004	127.1204862	1	Green2
PI407170	KOR	37.2840004	127.1204862	1	Green2
PI407171	KOR	37.2840004	127.1204862	1	Green2
PI407172	KOR	37.234169	127.2841682	1	Green2
PI407174	KOR	37.204679	127.4425791	1	Green2
PI407175	KOR	37.204679	127.4425791	1	Green2
PI407176	KOR	37.204679	127.4425791	1	Green2
PI407177	KOR	37.204679	127.4425791	1	Green2
PI407178	KOR	37.35	127.4425791	1	Green2
PI407179	KOR	37.7833328	127.1166649	1	Green2
PI407180	KOR	37.7833328	127.1166649	1	Green2
PI407181	KOR	37.66	127.315	1	Green2
PI407182	KOR	37.744	127.425	1	Green2
PI407183	KOR	37.744	127.425	1	Green2
PI407184	KOR	37.2391124	127.0191689	1	Green2
PI407185	KOR	37.1171747	127.0593852	1	Green2
PI407186	KOR	37.1171747	127.0593852	1	Green2
PI407187	KOR	37.1171747	127.0593852	1	Green2
PI407188	KOR	37.2072	126.989	1	Green2
PI407190	KOR	37.2072	126.989	1	Green2
PI407192	KOR	37.82	127.715	1	Green2
PI407193	KOR	37.75	127.795	1	Green2
PI407194	KOR	37.8182049	127.7499924	1	Green2
PI407195	KOR	37.68	127.88	1	Green2
PI407196	KOR	37.68	127.88	1	Green2
PI407198	KOR	37.62	127.82	1	Green2
PI407199	KOR	37.62	127.82	1	Green2
PI407200	KOR	37.4387489	127.9833336	1	Green2
PI407201	KOR	37.4387489	127.9833336	1	Green2
PI407202	KOR	37.5018208	127.9833336	1	Green2
PI407203	KOR	37.2945847	127.918	1	Green2
PI407204	KOR	37.2945847	127.918	1	Green2
PI407205	KOR	37.279	127.909	1	Green2
PI407206	KOR	37.18	127.89	1	Green2
PI407207	KOR	37.158	127.886	1	Green2
PI407208	KOR	37.158	127.886	1	Green2

PI407209	KOR	37.11	127.878	1	Green2
PI407211	KOR	37.08	127.89	1	Green2
PI407212	KOR	37.08	127.89	1	Green2
PI407213	KOR	37.08	127.89	1	Green2
PI407214	KOR	37.0455	127.94555	1	Green2
PI407215	KOR	37.0455	127.94555	1	Green2
PI407216	KOR	36.96	127.85	1	Green2
PI407217	KOR	36.96	127.85	1	Green2
PI407219	KOR	36.94535	127.7396	1	Green2
PI407220	KOR	36.94535	127.7396	1	Green2
PI407222	KOR	36.855	127.63	1	Green2
PI407224	KOR	36.759935	127.550753	1	Green2
PI407226	KOR	36.586863	127.4256134	1	Green2
PI407228	KOR	36.586863	127.4256134	1	Green2
PI407229	KOR	36.5	127.237	1	Green2
PI407231	KOR	36.5	127.237	1	Green2
PI407232	KOR	36.5	127.2	1	Green2
PI407233	KOR	36.62	127.291	1	Green2
PI407234	KOR	36.62	127.291	1	Green2
PI407235	KOR	36.62	127.291	1	Green2
PI407237	KOR	36.65	127.27	1	Green2
PI407238	KOR	36.68	127.2	1	Green2
PI407239	KOR	36.7608133	127.1621736	1	Green2
PI407240	KOR	35.6014929	128.7488937	1	Green2
PI407241	KOR	35.6014929	128.7488937	1	Green2
PI407242	KOR	35.612589	128.7499924	1	Green2
PI407243	KOR	35.612589	128.7499924	2	Red
PI407244	KOR	35.612589	128.7499924	1	Green2
PI407246	KOR	35.6846917	128.7499924	1	Green2
PI407247	KOR	35.6846917	128.7499924	1	Green2
PI407248	KOR	35.6846917	128.7499924	1	Green2
PI407249	KOR	35.666666	128.6616352	2	Red
PI407250	KOR	35.666666	128.6174566	1	Green2
PI407253	KOR	35.5127464	128.7655872	1	Green2
PI407254	KOR	35.5188286	128.7800654	1	Green2
PI407255	KOR	35.5188286	128.7800654	1	Green2
PI407256	KOR	35.5570679	128.8268231	1	Green2
PI407257	KOR	35.5570679	128.8268231	1	Green2
PI407258	KOR	35.5450654	128.75	2	Red
PI407259	KOR	35.5450654	128.75	1	Green2
PI407261	KOR	35.582	128.513	1	Green2
PI407262	KOR	35.5085039	128.492214	1	Green2
PI407263	KOR	35.5085039	128.492214	1	Green2
PI407265	KOR	35.44	128.56	1	Green2
PI407266	KOR	35.3980285	128.6253027	1	Green2
PI407267	KOR	35.3980285	128.6253027	1	Green2

PI407268	KOR	35.3980285	128.6253027	1	Green2
PI407269	KOR	35.4490143	128.6876514	1	Green2
PI407270	KOR	35.4490143	128.6876514	1	Green2
PI407271	KOR	35.5577865	126.8704605	1	Green2
PI407272	KOR	35.523844	126.8968964	1	Green2
PI407273	KOR	35.523844	126.8968964	1	Green2
PI407275	KOR	37.4288883	126.9891701	1	Green2
PI407276	KOR	37.4288883	126.9891701	1	Green2
PI407277	KOR	37.4288883	126.9891701	1	Green2
PI407278	KOR	37.5667	127.2274761	1	Green2
PI407285	JPN	35.5869684	139.3450928	2	Red
PI407286	JPN	35.5869684	139.3450928	2	Red
PI407288	CHN	43.5072231	124.8122215	1	Green2
PI407289	CHN	43.5072231	124.8122215	1	Green2
PI407290	CHN	43.5072231	124.8122215	1	Green2
PI407291	CHN	43.5072231	124.8122215	1	Green2
PI407292	CHN	43.848421	125.309283	1	Green2
PI407293	CHN	43.848421	125.309283	1	Green2
PI407294	CHN	43.848421	125.309283	1	Green2
PI407295	CHN	43.848421	125.309283	1	Green2
PI407296	CHN	41.64	123.483611	1	Green2
PI407297	CHN	41.64	123.483611	1	Green2
PI407299	CHN	41.64	123.483611	1	Green2
PI407300	CHN	32.06	118.85	1	Green2
PI407301	CHN	32.06	118.85	1	Green2
PI407302	CHN	32.06	118.85	1	Green2
PI407303	CHN	32.06	118.85	1	Green2
PI407304	CHN	31.0177044	121.4086113	1	Green2
PI407305	CHN	31.0177044	121.4086113	1	Green2
PI407306	CHN	31.0177044	121.4086113	1	Green2
PI407307	CHN	31.148818	121.80095	1	Green2
PI407308	KOR	37.2841644	127.0191689	1	Green2
PI407310	KOR	37.08	127.42	1	Green2
PI407311	KOR	37.012	127.32	1	Green2
PI407312	KOR	36.867	127.5407	1	Green2
PI407313	KOR	36.9471	128.0709	1	Green2
PI407314	KOR	36.9935	128.3175	1	Green2
PI407315	KOR	36.933	128.98	1	Green2
PI407317	KOR	36.678	127.212	1	Green2
PI407319	KOR	36.471177	127.702817	1	Green2
PI407320	KOR	37	127.583	3	Blue
PI407321	KOR	36.62	127.37	1	Green2
PI407322	KOR	36.33	127.5348	1	Green2
PI423991	RUS	52.9775503	127.3620871	3	Blue
PI423992	RUS	52.9775503	127.3620871	3	Blue
PI423993	RUS	52.9775503	127.3620871	3	Blue

PI423994	RUS	52.9775503	127.3620871	3	Blue
PI423995	RUS	52.9775503	127.3620871	3	Blue
PI423996	RUS	52.9775503	127.3620871	3	Blue
PI423997	RUS	52.9775503	127.3620871	3	Blue
PI423998	RUS	52.9775503	127.3620871	3	Blue
PI424000	RUS	52.9775503	127.3620871	3	Blue
PI424001	RUS	52.9775503	127.3620871	1	Green2
PI424002	RUS	52.9775503	127.3620871	3	Blue
PI424003	RUS	52.9775503	127.3620871	3	Blue
PI424007	KOR	37.208889	126.8166695	1	Green2
PI424009	KOR	37.208889	126.8166695	1	Green2
PI424011	KOR	37.2288249	126.9695091	1	Green2
PI424012	KOR	37.2288249	126.9695091	1	Green2
PI424014	KOR	37.2841644	127.0191689	1	Green2
PI424015	KOR	37.875	127.025	1	Green2
PI424016	KOR	38.0929404	127.0755861	1	Green2
PI424018	KOR	38.0929404	127.0755861	1	Green2
PI424019	KOR	38.135	127.02	1	Green2
PI424023	KOR	38.05	127.26	1	Green2
PI424026	KOR	38.05	127.26	1	Green2
PI424029	KOR	38.07	127.3	1	Green2
PI424030	KOR	37.8236122	127.5141678	1	Green2
PI424032	KOR	37.898	126.977	1	Green2
PI424033	KOR	37.899	126.977	1	Green2
PI424035	KOR	37.899	126.977	1	Green2
PI424036	KOR	37.7499962	127.0833321	1	Green2
PI424037	KOR	37.899	126.977	1	Green2
PI424040	KOR	37.4865	127.654	1	Green2
PI424041	KOR	37.52	127.44	1	Green2
PI424042	KOR	37.5436115	127.3275032	1	Green2
PI424044	KOR	37.55	127.255	1	Green2
PI424045	KOR	37.55	127.255	1	Green2
PI424047	KOR	37.612506	127.2141685	1	Green2
PI424048	KOR	37.612506	127.2141685	1	Green2
PI424049	KOR	37.65	127.19	1	Green2
PI424050	KOR	37.6366673	127.2141685	1	Green2
PI424052	KOR	37.6366673	127.2141685	1	Green2
PI424053	KOR	37.612506	127.2141685	1	Green2
PI424055	KOR	37.65	127.188	1	Green2
PI424056	KOR	38.12868	127.348	1	Green2
PI424057	KOR	38.138	127.3	1	Green2
PI424058	KOR	38.18	127.355	1	Green2
PI424060	KOR	38.15825	127.414767	1	Green2
PI424062	KOR	37.94	127.748	1	Green2
PI424063	KOR	38.15825	127.414767	1	Green2
PI424064	KOR	38.082397	128.052565	1	Green2

PI424065	KOR	38.082397	128.052565	1	Green2
PI424066	KOR	38.12	128.22	1	Green2
PI424067	KOR	38.12	128.22	1	Green2
PI424068	KOR	38.0655575	128.1730652	1	Green2
PI424069	KOR	37.696952	127.888683	1	Green2
PI424071	KOR	37.643	127.8	1	Green2
PI424073	KOR	37.493	128.023	1	Green2
PI424074	KOR	37.481	128.033	1	Green2
PI424077	KOR	37.5013	128.464	1	Green2
PI424079	KOR	37.49	128.872	1	Green2
PI424082	KOR	37.26	128.42	1	Green2
PI424084	KOR	37.177	128.3903099	1	Green2
PI424086	KOR	37.17	128.27	1	Green2
PI424087	KOR	37.133333	128.216667	1	Green2
PI424088	KOR	37.133333	128.216667	1	Green2
PI424089	KOR	36.934	128.181	1	Green2
PI424090	KOR	37.133333	128.216667	1	Green2
PI424093	KOR	36.985	128.362778	1	Green2
PI424095	KOR	36.2285	127.9105	1	Green2
PI424096	KOR	36.365	127.187	1	Green2
PI424097	KOR	36.11681	128.004029	1	Green2
PI424098	KOR	36.18128	128.1245	1	Green2
PI424101	KOR	36.3385	128.1305	1	Green2
PI424104	KOR	36.664	128.12	1	Green2
PI424105	KOR	36.586148	128.186797	1	Green2
PI424108	KOR	36.685	127.7	1	Green2
PI424110	KOR	35.77176	128.810085	1	Green2
PI424111	KOR	36.53	129.047	1	Green2
PI424112	KOR	36.416666	129.0833282	1	Green2
PI424113	KOR	36.405	129.172	1	Green2
PI424117	KOR	35.917	128.999	1	Green2
PI424118	KOR	35.833334	129.2499924	1	Green2
PI424119	KOR	35.990911	128.825511	1	Green2
PI424120	KOR	35.71	129.21	1	Green2
PI424121	KOR	34.9727745	128.3236237	1	Green2
PI424122	KOR	34.969722	128.349722	2	Red
PI424125	KOR	35.4894	126.9017	1	Green2
PI424126	KOR	35.4894	126.9017	1	Green2
PI424129	KOR	35.4894	126.9017	1	Green2
PI424130	KOR	35.4894	126.9017	1	Green2
PI447004	CHN	42.4977549	126.8299142	1	Green2
PI458535	CHN	48.2666683	126.6000023	3	Blue
PI458536	CHN	46.8641509	126.8548431	1	Green2
PI464867	CHN	50.2128163	126.8155901	3	Blue
PI464926	CHN	42.7228358	124.3313446	3	Blue
PI464928	CHN	41.7122899	124.9085886	3	Blue

PI468916	CHN	41.2013855	122.3415871	1	Green2
PI479767	CHN	48.4760799	127.9719961	1	Green2
PI483461	CHN	41.5352589	117.5571441	1	Green2
PI483463	CHN	38.7843764	113.4195597	1	Green2
PI483465	CHN	34.8806704	110.0100476	1	Green2
PI483466	CHN	36.2383803	116.8414652	1	Green2
PI483467	CHN	33.5090963	115.2268221	1	Green2
PI486220	JPN	35.1166668	138.9166718	2	Red
PI487428	JPN	39.7	141.2	2	Red
PI487429	JPN	35.5	139.5	2	Red
PI487430	JPN	42.6	142.1	2	Red
PI487431	JPN	31.2	130.6	2	Red
PI504286	KOR	36.810833	127.794722	1	Green2
PI504289	JPN	39.33333333	141	2	Red
PI507580	JPN	35.9999962	138.9999924	2	Red
PI507581	JPN	40.6333332	140.6000061	2	Red
PI507582	JPN	40.6836109	141.3597107	2	Red
PI507583	JPN	40.6836109	141.3597107	2	Red
PI507584	JPN	40.6836109	141.3597107	2	Red
PI507585	JPN	40.6836109	141.3597107	2	Red
PI507586	JPN	40.6836109	141.3597107	2	Red
PI507587	JPN	40.6836109	141.3597107	2	Red
PI507588	JPN	40.6	141.316667	2	Red
PI507589	JPN	39.483333	141.316667	2	Red
PI507591	JPN	39.700069	140.730588	2	Red
PI507592	JPN	39.700069	140.730588	2	Red
PI507593	JPN	37.2000008	140.3166656	2	Red
PI507595	JPN	37.009721	138.650564	2	Red
PI507596	JPN	38.1833324	139.4333344	2	Red
PI507597	JPN	38.0499992	139.4166718	2	Red
PI507599	JPN	35.8578156	140.3036074	2	Red
PI507602	JPN	35.7770593	140.7415712	2	Red
PI507603	JPN	35.7770593	140.7415712	2	Red
PI507604	JPN	35.7770593	140.7415712	2	Red
PI507605	JPN	36.3666668	140.4833374	2	Red
PI507606	JPN	36.3166676	139.5833282	2	Red
PI507607	JPN	36.3166676	139.5833282	2	Red
PI507608	JPN	36.7166672	139.6833344	2	Red
PI507609	JPN	36.5499992	139.7333374	2	Red
PI507611	JPN	36.5499992	139.7333374	2	Red
PI507612	JPN	36.5499992	139.7333374	2	Red
PI507613	JPN	36.3166676	139.1833344	2	Red
PI507615	JPN	36.3999977	138.2499924	2	Red
PI507616	JPN	36.3999977	138.2499924	2	Red
PI507617	JPN	40.0605545	124.5577812	2	Red
PI507620	JPN	36.6364681	137.9590988	2	Red

PI507621	JPN	36.6000023	138.0333405	2	Red
PI507622	JPN	36.6499977	138.3166656	2	Red
PI507623	JPN	35.7263889	139.4838943	2	Red
PI507624	JPN	35.4833374	137.4999924	2	Red
PI507625	JPN	36	136.25	2	Red
PI507627	JPN	34.9354819	137.2357177	2	Red
PI507628	JPN	34.7666683	137.3833313	2	Red
PI507629	JPN	35.2999992	138.9333344	2	Red
PI507631	JPN	34.5333328	135.9499969	2	Red
PI507632	JPN	34.4999962	135.8000031	2	Red
PI507633	JPN	35.2408484	135.4577486	2	Red
PI507634	JPN	34.7999992	134.9833374	2	Red
PI507635	JPN	34.9999962	134.9999924	2	Red
PI507636	JPN	34.666666	135.1416703	2	Red
PI507637	JPN	34.666666	135.1166687	2	Red
PI507638	JPN	34.7908287	134.8500061	2	Red
PI507640	JPN	34.8166676	135.4166718	2	Red
PI507641	JPN	35.3999977	134.7666626	2	Red
PI507643	JPN	33.9833374	132.7833405	2	Red
PI507644	JPN	33.8363876	132.753067	2	Red
PI507645	JPN	34.0666676	134.4499969	2	Red
PI507646	JPN	34.4833374	133.3666687	2	Red
PI507647	JPN	34.9999962	133.9999924	2	Red
PI507649	JPN	35	133.9166718	2	Red
PI507650	JPN	35.4999962	134.2333374	2	Red
PI507651	JPN	34.166666	131.4833374	2	Red
PI507652	JPN	34.1000023	131.3999939	2	Red
PI507653	JPN	33.8666668	130.7499924	2	Red
PI507654	JPN	33.2499962	130.3000031	2	Red
PI507657	JPN	31.9499998	130.7166672	2	Red
PI507660	JPN	31.499999	130.4166718	2	Red
PI507661	JPN	31.9499998	130.7166672	2	Red
PI507662	JPN	31.333333	130.9333344	2	Red
PI507663	JPN	31.6166649	130.3999939	2	Red
PI507664	JPN	32.6671247	130.6933593	2	Red
PI507666	JPN	32.8833332	131.1000061	2	Red
PI507667	JPN	32.6671247	130.6933593	2	Red
PI507668	JPN	32.6671247	130.6933593	2	Red
PI507669	JPN	32.6671247	130.6933593	2	Red
PI507722	RUS	52.9775503	127.3620871	3	Blue
PI507728	RUS	52.9775503	127.3620871	3	Blue
PI507730	RUS	52.9775503	127.3620871	3	Blue
PI507805	RUS	44.9999962	134.9999924	1	Green2
PI507847	RUS	52.9775503	127.3620871	3	Blue
PI508060	JPN	42.583334	142.1333313	2	Red
PI508063	JPN	42.583334	142.1333313	2	Red

PI508064	JPN	42.3669254	142.4114771	2	Red
PI508067	JPN	42.3669254	142.4114771	2	Red
PI508069	JPN	42.3669254	142.4114771	2	Red
PI514674	JPN	42.7550795	142.7453613	2	Red
PI522180	CHN	48.8191949	128.4075207	3	Blue
PI522181	CHN	48.2666683	126.6000023	3	Blue
PI522184	CHN	45.7333374	127.4500008	3	Blue
PI522193	RUS	52.9775503	127.3620871	3	Blue
PI522197	RUS	44.9999962	134.9999924	3	Blue
PI522199	RUS	44.9999962	134.9999924	3	Blue
PI522201	RUS	44.9999962	134.9999924	1	Green2
PI522202	RUS	44.9999962	134.9999924	3	Blue
PI522204	RUS	44.9999962	134.9999924	1	Green2
PI522206	RUS	44.9999962	134.9999924	3	Blue
PI522210	RUS	44.9999962	134.9999924	3	Blue
PI522215	RUS	44.9999962	134.9999924	1	Green2
PI522216	RUS	44.9999962	134.9999924	1	Green2
PI522217	RUS	44.9999962	134.9999924	1	Green2
PI522218	RUS	44.9999962	134.9999924	3	Blue
PI522221	RUS	44.9999962	134.9999924	1	Green2
PI522222	RUS	44.9999962	134.9999924	3	Blue
PI522223	RUS	44.9999962	134.9999924	3	Blue
PI522226	RUS	44.9999962	134.9999924	3	Blue
PI522227	RUS	44.9999962	134.9999924	3	Blue
PI522228	RUS	44.9999962	134.9999924	3	Blue
PI522229	RUS	44.9999962	134.9999924	1	Green2
PI522232	RUS	44.9999962	134.9999924	3	Blue
PI522234	RUS	44.9999962	134.9999924	3	Blue
PI532449	CHN	44	125	1	Green2
PI532450	CHN	42	126	1	Green2
PI532451	CHN	41	126	1	Green2
PI549032	CHN	40.5476498	124.0656356	1	Green2
PI549033	CHN	40.5476498	124.0656356	3	Blue
PI549034	CHN	40.5476498	124.0656356	1	Green2
PI549036	CHN	40.5476498	124.0656356	1	Green2
PI549037	CHN	40.5476498	124.0656356	1	Green2
PI549039	CHN	40.7142354	125.0417329	1	Green2
PI549046	CHN	37.5318223	107.3972706	1	Green2
PI549047	CHN	40.2215443	116.4283296	1	Green2
PI549048	CHN	40.1873243	116.1983376	1	Green2
PI562531	KOR	37.23333333	126.9333333	1	Green2
PI562532	KOR	37.23333333	126.9333333	1	Green2
PI562533	KOR	37.23333333	126.9333333	1	Green2
PI562534	KOR	37.23333333	126.9333333	1	Green2
PI562535	KOR	37.23333333	126.9333333	1	Green2
PI562536	KOR	37.23333333	126.9333333	1	Green2

PI562537	KOR	37.23333333	126.9333333	1	Green2
PI562538	KOR	37.23333333	126.9333333	1	Green2
PI562539	KOR	37.23333333	126.9333333	1	Green2
PI562540	KOR	37.23333333	126.9333333	1	Green2
PI562541	KOR	37.23333333	126.9333333	1	Green2
PI562542	KOR	37.23333333	126.9333333	1	Green2
PI562543	KOR	36.85	126.9333333	1	Green2
PI562544	KOR	36.85	126.9333333	1	Green2
PI562545	KOR	36.85	126.9333333	1	Green2
PI562546	KOR	36.85	126.9333333	1	Green2
PI562547	KOR	36.56666667	126.6833333	1	Green2
PI562548	KOR	36.56666667	126.6833333	1	Green2
PI562549	KOR	36.56666667	126.6833333	1	Green2
PI562550	KOR	36.56666667	126.6833333	1	Green2
PI562551	KOR	36.56666667	126.6833333	1	Green2
PI562552	KOR	36.56666667	126.6833333	1	Green2
PI562553	KOR	36.18333333	126.5666667	1	Green2
PI562554	KOR	36.18333333	126.5666667	1	Green2
PI562555	KOR	36.18333333	126.5666667	1	Green2
PI562556	KOR	35.81666667	127.1166667	1	Green2
PI562557	KOR	35.81666667	127.1166667	1	Green2
PI562558	KOR	35.81666667	127.1166667	1	Green2
PI562559	KOR	35.81666667	127.1166667	1	Green2
PI562561	KOR	35.81666667	127.1166667	1	Green2
PI562562	KOR	35.53333333	127.3333333	1	Green2
PI562563	KOR	35.53333333	127.3333333	1	Green2
PI562565	KOR	35.53333333	127.3333333	1	Green2
PI562566	KOR	35.53333333	127.3333333	1	Green2
PI562567	KOR	35.53333333	127.3333333	1	Green2
PI562568	KOR	35.53333333	127.3333333	1	Green2
PI567194	RUS	52.9775503	127.3620871	1	Green2
PI578336	RUS	52.9775503	127.3620871	3	Blue
PI578337	RUS	52.9775503	127.3620871	1	Green2
PI578341	RUS	48.4969043	135.1323167	3	Blue
PI578343	RUS	48.4969043	135.1323167	3	Blue
PI578345	RUS	48.4969043	135.1323167	3	Blue
PI593983	JPN	42.872776	142.440567	2	Red

Table S2. Significant hits from environmental association results. Position in gene and Minor Allele frequency (MAF) estimates from Song *et al.* 2013.

Trait	SNP	Pval	Significant in other Env Associations	Position in gene	MAF in Landrace	MAF in Elite (Song et al. 2013)	MAF in G. Soja (Song et al. 2013)
Altitude	BARC_1.01_Gm_20_4619978_A_G	1.10E-06	6	(non-genic)	0.469	0.182	0.12
Altitude	BARC_1.01_Gm_19_567731_A_G	4.93E-06	6	(non-genic)	0.147	0.074	0.112
Altitude	BARC_1.01_Gm_14_4649711_A_G	7.70E-06	0	(non-genic)	0	0	0.187
Annual_Precipitation	BARC_1.01_Gm_15_12227854_G_A	1.48E-05	2	Intron	0.402	0.188	0.052
Annual_Precipitation	BARC_1.01_Gm_17_8010009_A_C	2.67E-05	0	(non-genic)	0.212	0.484	0.158
Annual_Precipitation	BARC_1.01_Gm_07_2890463_A_G	3.30E-05	0	(non-genic)	0.26	0.328	0.421
Bulk_Density_Subsoil	BARC_1.01_Gm_04_45514250_A_G	2.20E-06	1	Intron	0.437	0.078	0.197
Bulk_Density_Subsoil	BARC_1.01_Gm_17_38522278_G_T	5.29E-05	1	(non-genic)	0.048	0.242	0.258
Bulk_Density_Subsoil	BARC_1.01_Gm_20_35812683_A_G	5.57E-05	1	3UTR	0.187	0.179	0.299
Bulk_Density_Topsoil	BARC_1.01_Gm_04_45514250_A_G	2.20E-06	1	Intron	0.437	0.078	0.197
Bulk_Density_Topsoil	BARC_1.01_Gm_17_38522278_G_T	5.37E-05	1	(non-genic)	0.048	0.242	0.258
Bulk_Density_Topsoil	BARC_1.01_Gm_20_35812683_A_G	5.41E-05	1	3UTR	0.187	0.179	0.299
Cation_Exchange_Capacity_Subsoil	BARC_1.01_Gm_07_4921108_G_A	2.58E-05	1	Intron	0.479	0.297	0.484
Cation_Exchange_Capacity_Subsoi	BARC_1.01_Gm_07_41769344_G_T	3.33E-05	1	(non-genic)	0.021	0	0.221

l							
Cation_Exchange_Capacity_Subsoil	BARC_1.01_Gm_09_3232979_A_G	7.42E-05	0	CDS	0.253	0.311	0.134
Cation_Exchange_Capacity_Topsoil	BARC_1.01_Gm_07_4921108_G_A	3.80E-05	1	Intron	0.479	0.297	0.484
Cation_Exchange_Capacity_Topsoil	BARC_1.01_Gm_07_41769344_G_T	4.93E-05	1	(non-genic)	0.021	0	0.221
Cation_Exchange_Capacity_Topsoil	BARC_1.01_Gm_14_23750665_G_A	5.26E-05	4	(non-genic)	0	0	0.061
Isothermality	BARC_1.01_Gm_14_14468456_C_T	7.90E-06	0	(non-genic)	0.031	0	0.247
Isothermality	BARC_1.01_Gm_07_34000545_G_T	5.86E-05	0	(non-genic)	0.01	0	0.172
Isothermality	BARC_1.01_Gm_07_33356669_G_A	6.18E-05	0	(non-genic)	0.01	0	0.178
Max_Temp_April	BARC_1.01_Gm_20_968323_A_G	5.01E-07	10	(non-genic)	0.104	0	0
Max_Temp_April	BARC_1.01_Gm_07_40528989_A_G	4.86E-05	0	(non-genic)	0.219	0.021	0.175
Max_Temp_April	BARC_1.01_Gm_15_7608425_T_C	5.51E-05	1	3UTR	0.284	0.354	0.5
Max_Temp_August	BARC_1.01_Gm_16_1552499_A_G	7.30E-06	10	(non-genic)	0.031	0.295	0.151
Max_Temp_August	BARC_1.01_Gm_20_968323_A_G	1.19E-05	10	(non-genic)	0.104	0	0
Max_Temp_August	BARC_1.01_Gm_19_567731_A_G	1.33E-05	6	(non-genic)	0.147	0.074	0.112
Max_Temp_December	BARC_1.01_Gm_04_8023658_C_T	2.84E-05	8	CDS	0.333	0.205	0.116
Max_Temp_December	BARC_1.01_Gm_05_40413855_G_A	2.96E-05	2	3UTR	0.3	0.191	0.362

Max_Temp_December	BARC_1.01_Gm_04_8017920_T_C	3.32E-05	4	CDS	0.33	0.205	0.105
Max_Temp_February	BARC_1.01_Gm_02_7980013_G_A	4.97E-06	3	(non-genic)	0.422	0.297	0.214
Max_Temp_February	BARC_1.01_Gm_02_7974982_C_T	7.63E-05	0	(non-genic)	0.41	0.297	0.208
Max_Temp_February	BARC_1.01_Gm_09_43103646_G_A	7.98E-05	0	CDS	0.276	0.253	0.166
Max_Temp_January	BARC_1.01_Gm_05_40413855_G_A	2.00E-06	2	3UTR	0.3	0.191	0.362
Max_Temp_January	BARC_1.01_Gm_02_7980013_G_A	5.82E-06	3	(non-genic)	0.422	0.297	0.214
Max_Temp_January	BARC_1.01_Gm_13_22126286_G_T	9.56E-06	11	(non-genic)	0.26	0.239	0.362
Max_Temp_July	BARC_1.01_Gm_16_1552499_A_G	1.18E-06	10	(non-genic)	0.031	0.295	0.151
Max_Temp_July	BARC_1.01_Gm_20_968323_A_G	1.77E-05	10	(non-genic)	0.104	0	0
Max_Temp_July	BARC_1.01_Gm_10_22209844_G_A	2.94E-05	0	(non-genic)	0.011	0	0.071
Max_Temp_June	BARC_1.01_Gm_20_968323_A_G	3.38E-07	10	(non-genic)	0.104	0	0
Max_Temp_June	BARC_1.01_Gm_16_1552499_A_G	4.28E-06	10	(non-genic)	0.031	0.295	0.151
Max_Temp_June	BARC_1.01_Gm_09_5664883_A_G	6.54E-06	1	(non-genic)	0.141	0.028	0.29
Max_Temp_March	BARC_1.01_Gm_01_16610088_T_C	3.16E-05	0	(non-genic)	0.156	0	0.264
Max_Temp_March	BARC_1.01_scaffold_23_881897_T_C	4.46E-05	0	(non-genic)	0.147	0	0.093
Max_Temp_March	BARC_1.01_Gm_08_22586252_C_A	5.16E-05	1	(non-genic)	0.479	0.3	0.053
Max_Temp_May	BARC_1.01_Gm_20_968323_A_G	1.18E-06	10	(non-genic)	0.104	0	0
Max_Temp_May	BARC_1.01_Gm_09_5664883_A_G	5.80E-05	1	(non-genic)	0.141	0.028	0.29

Max_Temp_May	BARC_1.01_Gm_15_7608425_T_C	7.29E-05	1	3UTR	0.284	0.354	0.5
Max_Temp_November	BARC_1.01_Gm_14_46604963_T_C	8.43E-05	0	(non-genic)	0.358	0.094	0.055
Max_Temp_November	BARC_1.01_Gm_08_6718151_A_G	1.17E-04	0	(non-genic)	0	0	0.434
Max_Temp_November	BARC_1.01_Gm_04_8023658_C_T	1.18E-04	8	CDS	0.333	0.205	0.116
Max_Temp_October	BARC_1.01_Gm_20_4619978_A_G	6.69E-06	6	(non-genic)	0.469	0.182	0.12
Max_Temp_October	BARC_1.01_Gm_16_434918_T_C	6.12E-05	1	Intron	0.305	0.468	0.143
Max_Temp_October	BARC_1.01_Gm_08_22586252_C_A	6.20E-05	1	(non-genic)	0.479	0.3	0.053
Max_Temp_September	BARC_1.01_Gm_20_4619978_A_G	1.10E-05	6	(non-genic)	0.469	0.182	0.12
Max_Temp_September	BARC_1.01_Gm_19_567731_A_G	2.68E-05	6	(non-genic)	0.147	0.074	0.112
Max_Temp_September	BARC_1.01_Gm_01_26114693_T_C	3.96E-05	10	(non-genic)	0.01	0	0.216
Max_Temp_Warmest_Month	BARC_1.01_Gm_16_1552499_A_G	7.61E-07	10	(non-genic)	0.031	0.295	0.151
Max_Temp_Warmest_Month	BARC_1.01_Gm_20_968323_A_G	1.93E-06	10	(non-genic)	0.104	0	0
Max_Temp_Warmest_Month	BARC_1.01_Gm_19_567731_A_G	3.46E-05	6	(non-genic)	0.147	0.074	0.112
Mean_Annual_Temperature	BARC_1.01_Gm_15_24786409_C_T	7.94E-06	9	(non-genic)	0.085	0.211	0.073
Mean_Annual_Temperature	BARC_1.01_Gm_01_26114693_T_C	1.34E-05	10	(non-genic)	0.01	0	0.216
Mean_Annual_Temperature	BARC_1.01_Gm_20_4619978_A_G	6.23E-05	6	(non-genic)	0.469	0.182	0.12
Mean_Diurnal_Range	BARC_1.01_Gm_10_49584311_T_C	2.59E-05	0	(non-genic)	0.453	0.408	0.109
Mean_Diurnal_Range	BARC_1.01_Gm_18_11964908_T_C	3.21E-05	0	(non-genic)	0.328	0.234	0.273

Mean_Diurnal_Range	BARC_1.01_Gm_05_4252974_A_G	8.23E-05	1	(non-genic)	0	0	0.095
Mean_Precipitation_April	BARC_1.01_Gm_02_49540930_T_C	5.82E-06	1	Intron	0.324	0.094	0.253
Mean_Precipitation_April	BARC_1.01_Gm_13_23459258_C_T	7.17E-06	2	(non-genic)	0.404	0.167	0.105
Mean_Precipitation_April	BARC_1.01_Gm_01_43461285_A_G	1.77E-05	0	(non-genic)	0.469	0.365	0.081
Mean_Precipitation_August	BARC_1.01_Gm_09_38979856_T_C	2.06E-05	1	(non-genic)	0.115	0.089	0.383
Mean_Precipitation_August	BARC_1.01_Gm_04_41672171_T_C	5.87E-05	0	CDS	0.198	0.417	0.495
Mean_Precipitation_August	BARC_1.01_Gm_04_18102418_T_C	1.89E-04	0	(non-genic)	0.01	0.411	0.271
Mean_Precipitation_December	BARC_1.01_Gm_03_37858728_G_A	4.69E-08	0	(non-genic)	0.394	0.247	0.08
Mean_Precipitation_December	BARC_1.01_Gm_15_11564048_G_A	1.07E-07	0	(non-genic)	0.375	0.063	0.112
Mean_Precipitation_December	BARC_1.01_Gm_16_30075422_T_C	1.34E-06	2	Intron	0.311	0.332	0.14
Mean_Precipitation_February	BARC_1.01_Gm_02_36762041_T_G	1.96E-08	0	(non-genic)	0.332	0.276	0.447
Mean_Precipitation_February	BARC_1.01_Gm_02_41646843_T_C	2.72E-08	2	(non-genic)	0.042	0.01	0.492
Mean_Precipitation_February	BARC_1.01_Gm_02_41663747_A_G	5.02E-08	2	(non-genic)	0.042	0.01	0.386
Mean_Precipitation_January	BARC_1.01_Gm_16_30075422_T_C	8.31E-08	2	Intron	0.311	0.332	0.14
Mean_Precipitation_January	BARC_1.01_Gm_02_41646843_T_C	3.34E-07	2	(non-genic)	0.042	0.01	0.492
Mean_Precipitation_January	BARC_1.01_Gm_02_41663747_A_G	5.70E-07	2	(non-genic)	0.042	0.01	0.386
Mean_Precipitation_July	BARC_1.01_Gm_08_2254106_G_A	2.00E-07	1	CDS	0.224	0.01	0.332
Mean_Precipitation_July	BARC_1.01_Gm_09_38979856_T_C	1.70E-06	1	(non-genic)	0.115	0.089	0.383

Mean_Precipitation_July	BARC_1.01_Gm_18_8791607_C_T	2.75E-06	0	(non-genic)	0.182	0.068	0.137
Mean_Precipitation_June	BARC_1.01_Gm_13_23459258_C_T	3.17E-07	2	(non-genic)	0.404	0.167	0.105
Mean_Precipitation_June	BARC_1.01_Gm_02_49540930_T_C	7.15E-05	1	Intron	0.324	0.094	0.253
Mean_Precipitation_June	BARC_1.01_Gm_15_6765154_C_T	1.22E-04	0	(non-genic)	0.245	0.5	0.327
Mean_Precipitation_March	BARC_1.01_Gm_12_277889_G_A	1.67E-07	2	CDS	0.302	0.276	0.253
Mean_Precipitation_March	BARC_1.01_Gm_08_44927121_T_C	2.29E-07	0	CDS	0.042	0.01	0.071
Mean_Precipitation_March	BARC_1.01_Gm_14_48129511_T_C	7.60E-06	0	(non-genic)	0.365	0.279	0.407
Mean_Precipitation_May	BARC_1.01_Gm_13_23459258_C_T	2.27E-06	2	(non-genic)	0.404	0.167	0.105
Mean_Precipitation_May	BARC_1.01_Gm_17_37308670_C_T	4.64E-06	0	(non-genic)	0.126	0.311	0.089
Mean_Precipitation_May	BARC_1.01_Gm_12_277889_G_A	6.40E-06	2	CDS	0.302	0.276	0.253
Mean_Precipitation_November	BARC_1.01_Gm_07_7709976_G_T	2.28E-07	0	Intron	0.052	0.375	0.199
Mean_Precipitation_November	BARC_1.01_Gm_10_34480265_C_A	2.05E-06	1	(non-genic)	0.104	0.01	0.111
Mean_Precipitation_November	BARC_1.01_Gm_20_36582013_T_C	5.56E-06	0	(non-genic)	0	0	0
Mean_Precipitation_October	BARC_1.01_Gm_16_30386356_A_C	2.28E-08	0	(non-genic)	0.484	0.043	0.383
Mean_Precipitation_October	BARC_1.01_Gm_15_12227854_G_A	2.08E-07	2	Intron	0.402	0.188	0.052
Mean_Precipitation_October	BARC_1.01_Gm_19_40490186_A_G	4.95E-07	0	(non-genic)	0.063	0.179	0.135
Mean_Precipitation_September	BARC_1.01_Gm_15_12227854_G_A	1.08E-06	2	Intron	0.402	0.188	0.052
Mean_Precipitation_September	BARC_1.01_Gm_12_277889_G_A	1.23E-05	2	CDS	0.302	0.276	0.253

Mean_Precipitation_September	BARC_1.01_Gm_03_4782127_T_C	6.10E-05	0	(non-genic)	0.479	0.312	0.344
Mean_Temp_April	BARC_1.01_Gm_20_968323_A_G	1.82E-05	10	(non-genic)	0.104	0	0
Mean_Temp_April	BARC_1.01_Gm_20_549607_G_A	3.17E-05	0	Intron	0.451	0.105	0.237
Mean_Temp_April	BARC_1.01_Gm_15_24786409_C_T	3.57E-05	9	(non-genic)	0.085	0.211	0.073
Mean_Temp_August	BARC_1.01_Gm_19_567731_A_G	1.62E-05	6	(non-genic)	0.147	0.074	0.112
Mean_Temp_August	BARC_1.01_Gm_16_1552499_A_G	2.31E-05	10	(non-genic)	0.031	0.295	0.151
Mean_Temp_August	BARC_1.01_Gm_07_7582760_T_C	4.46E-05	0	(non-genic)	0.438	0.453	0.418
Mean_Temp_Coldest_Quarter	BARC_1.01_Gm_13_22126286_G_T	7.26E-07	11	(non-genic)	0.26	0.239	0.362
Mean_Temp_Coldest_Quarter	BARC_1.01_Gm_05_40413855_G_A	6.61E-06	2	3UTR	0.3	0.191	0.362
Mean_Temp_Coldest_Quarter	BARC_1.01_Gm_04_8023658_C_T	1.52E-05	8	CDS	0.333	0.205	0.116
Mean_Temp_December	BARC_1.01_Gm_15_47013300_T_C	3.33E-06	2	(non-genic)	0.299	0.126	0.214
Mean_Temp_December	BARC_1.01_Gm_13_22126286_G_T	9.05E-06	11	(non-genic)	0.26	0.239	0.362
Mean_Temp_December	BARC_1.01_Gm_04_8023658_C_T	1.47E-05	8	CDS	0.333	0.205	0.116
Mean_Temp_Driest_Quarter	BARC_1.01_Gm_02_7317092_T_C	8.29E-07	0	Intron	0.342	0.207	0.031
Mean_Temp_Driest_Quarter	BARC_1.01_Gm_12_32654107_G_T	8.73E-06	0	(non-genic)	0.094	0.073	0.136
Mean_Temp_Driest_Quarter	BARC_1.01_Gm_02_44737130_T_G	9.80E-06	0	(non-genic)	0.125	0.097	0.484
Mean_Temp_February	BARC_1.01_Gm_01_26114693_T_C	2.22E-05	10	(non-genic)	0.01	0	0.216
Mean_Temp_February	BARC_1.01_Gm_13_22126286_G_T	2.55E-05	11	(non-genic)	0.26	0.239	0.362

Mean_Temp_February	BARC_1.01_Gm_02_7980013_G_A	3.14E-05	3	(non-genic)	0.422	0.297	0.214
Mean_Temp_January	BARC_1.01_Gm_04_8023658_C_T	1.64E-05	8	CDS	0.333	0.205	0.116
Mean_Temp_January	BARC_1.01_Gm_13_22126286_G_T	2.32E-05	11	(non-genic)	0.26	0.239	0.362
Mean_Temp_January	BARC_1.01_Gm_04_8017920_T_C	2.52E-05	4	CDS	0.33	0.205	0.105
Mean_Temp_July	BARC_1.01_Gm_16_1552499_A_G	2.32E-06	10	(non-genic)	0.031	0.295	0.151
Mean_Temp_July	BARC_1.01_Gm_20_968323_A_G	7.78E-06	10	(non-genic)	0.104	0	0
Mean_Temp_July	BARC_1.01_Gm_15_24786409_C_T	2.60E-05	9	(non-genic)	0.085	0.211	0.073
Mean_Temp_June	BARC_1.01_Gm_16_1552499_A_G	1.53E-06	10	(non-genic)	0.031	0.295	0.151
Mean_Temp_June	BARC_1.01_Gm_20_968323_A_G	1.58E-06	10	(non-genic)	0.104	0	0
Mean_Temp_June	BARC_1.01_Gm_08_29906080_T_C	3.13E-05	1	(non-genic)	0.115	0.359	0.346
Mean_Temp_March	BARC_1.01_Gm_02_7980013_G_A	2.87E-05	3	(non-genic)	0.422	0.297	0.214
Mean_Temp_March	BARC_1.01_Gm_06_15569871_T_G	6.22E-05	0	(non-genic)	0.292	0.367	0.424
Mean_Temp_March	BARC_1.01_Gm_16_29518026_C_T	9.23E-05	0	(non-genic)	0	0	0.176
Mean_Temp_May	BARC_1.01_Gm_20_968323_A_G	1.02E-05	10	(non-genic)	0.104	0	0
Mean_Temp_May	BARC_1.01_Gm_16_1552499_A_G	2.68E-05	10	(non-genic)	0.031	0.295	0.151
Mean_Temp_May	BARC_1.01_Gm_08_29906080_T_C	3.29E-05	1	(non-genic)	0.115	0.359	0.346
Mean_Temp_November	BARC_1.01_Gm_15_47013300_T_C	1.57E-05	2	(non-genic)	0.299	0.126	0.214
Mean_Temp_November	BARC_1.01_Gm_04_8023658_C_T	5.29E-05	8	CDS	0.333	0.205	0.116

Mean_Temp_November	BARC_1.01_Gm_13_22126286_G_T	6.27E-05	11	(non-genic)	0.26	0.239	0.362
Mean_Temp_October	BARC_1.01_Gm_01_26114693_T_C	8.92E-06	10	(non-genic)	0.01	0	0.216
Mean_Temp_October	BARC_1.01_Gm_20_4619978_A_G	1.22E-05	6	(non-genic)	0.469	0.182	0.12
Mean_Temp_October	BARC_1.01_Gm_15_24786409_C_T	1.97E-05	9	(non-genic)	0.085	0.211	0.073
Mean_Temp_September	BARC_1.01_Gm_01_26114693_T_C	6.07E-06	10	(non-genic)	0.01	0	0.216
Mean_Temp_September	BARC_1.01_Gm_20_4619978_A_G	1.50E-05	6	(non-genic)	0.469	0.182	0.12
Mean_Temp_September	BARC_1.01_Gm_15_24786409_C_T	3.14E-05	9	(non-genic)	0.085	0.211	0.073
Mean_Temp_Warmest_Quarter	BARC_1.01_Gm_16_1552499_A_G	1.18E-06	10	(non-genic)	0.031	0.295	0.151
Mean_Temp_Warmest_Quarter	BARC_1.01_Gm_20_968323_A_G	1.24E-05	10	(non-genic)	0.104	0	0
Mean_Temp_Warmest_Quarter	BARC_1.01_Gm_15_24786409_C_T	1.98E-05	9	(non-genic)	0.085	0.211	0.073
Mean_Temperature_Wettest_Quarter	BARC_1.01_Gm_08_40882335_A_G	1.47E-06	0	(non-genic)	0.353	0.036	0.104
Mean_Temperature_Wettest_Quarter	BARC_1.01_Gm_13_31206278_G_A	3.79E-06	2	CDS	0.096	0.463	0.032
Mean_Temperature_Wettest_Quarter	BARC_1.01_Gm_08_40883682_C_T	6.78E-06	0	(non-genic)	0.342	0.036	0.104
Min_Temp_April	BARC_1.01_Gm_01_26114693_T_C	7.75E-06	10	(non-genic)	0.01	0	0.216
Min_Temp_April	BARC_1.01_Gm_07_39734469_T_C	2.90E-05	1	CDS	0	0	0.324
Min_Temp_April	BARC_1.01_Gm_15_24786409_C_T	3.32E-05	9	(non-genic)	0.085	0.211	0.073

Min_Temp_August	BARC_1.01_Gm_01_472038_C_A	1.31E-05	3	(non-genic)	0.221	0.442	0.234
Min_Temp_August	BARC_1.01_Gm_19_567731_A_G	2.89E-05	6	(non-genic)	0.147	0.074	0.112
Min_Temp_August	BARC_1.01_Gm_15_24786409_C_T	3.39E-05	9	(non-genic)	0.085	0.211	0.073
Min_Temp_Coldest_Month	BARC_1.01_Gm_04_8023658_C_T	5.19E-06	8	CDS	0.333	0.205	0.116
Min_Temp_Coldest_Month	BARC_1.01_Gm_13_22126286_G_T	5.46E-06	11	(non-genic)	0.26	0.239	0.362
Min_Temp_Coldest_Month	BARC_1.01_Gm_04_8017920_T_C	7.27E-06	4	CDS	0.33	0.205	0.105
Min_Temp_December	BARC_1.01_Gm_04_8023658_C_T	1.84E-06	8	CDS	0.333	0.205	0.116
Min_Temp_December	BARC_1.01_Gm_04_8017920_T_C	2.33E-06	4	CDS	0.33	0.205	0.105
Min_Temp_December	BARC_1.01_Gm_10_34480265_C_A	1.58E-05	1	(non-genic)	0.104	0.01	0.111
Min_Temp_February	BARC_1.01_Gm_13_22126286_G_T	2.41E-05	11	(non-genic)	0.26	0.239	0.362
Min_Temp_February	BARC_1.01_Gm_01_26114693_T_C	4.60E-05	10	(non-genic)	0.01	0	0.216
Min_Temp_February	BARC_1.01_Gm_04_6006337_T_C	4.69E-05	1	(non-genic)	0.161	0.347	0.061
Min_Temp_January	BARC_1.01_Gm_04_8023658_C_T	5.16E-06	8	CDS	0.333	0.205	0.116
Min_Temp_January	BARC_1.01_Gm_13_22126286_G_T	5.80E-06	11	(non-genic)	0.26	0.239	0.362
Min_Temp_January	BARC_1.01_Gm_04_8017920_T_C	7.20E-06	4	CDS	0.33	0.205	0.105
Min_Temp_July	BARC_1.01_Gm_01_472038_C_A	8.50E-06	3	(non-genic)	0.221	0.442	0.234
Min_Temp_July	BARC_1.01_Gm_16_1552499_A_G	1.21E-05	10	(non-genic)	0.031	0.295	0.151
Min_Temp_July	BARC_1.01_Gm_19_567731_A_G	2.14E-05	6	(non-genic)	0.147	0.074	0.112

Min_Temp_June	BARC_1.01_Gm_16_1552499_A_G	5.01E-06	10	(non-genic)	0.031	0.295	0.151
Min_Temp_June	BARC_1.01_Gm_01_472038_C_A	1.34E-05	3	(non-genic)	0.221	0.442	0.234
Min_Temp_June	BARC_1.01_Gm_15_24786409_C_T	1.78E-05	9	(non-genic)	0.085	0.211	0.073
Min_Temp_March	BARC_1.01_Gm_20_4619978_A_G	2.92E-05	6	(non-genic)	0.469	0.182	0.12
Min_Temp_March	BARC_1.01_Gm_16_434918_T_C	4.43E-05	1	Intron	0.305	0.468	0.143
Min_Temp_March	BARC_1.01_Gm_13_22126286_G_T	5.46E-05	11	(non-genic)	0.26	0.239	0.362
Min_Temp_May	BARC_1.01_Gm_15_24786409_C_T	1.09E-05	9	(non-genic)	0.085	0.211	0.073
Min_Temp_May	BARC_1.01_Gm_07_39734469_T_C	1.83E-05	1	CDS	0	0	0.324
Min_Temp_May	BARC_1.01_Gm_01_472038_C_A	3.09E-05	3	(non-genic)	0.221	0.442	0.234
Min_Temp_November	BARC_1.01_Gm_15_47013300_T_C	8.65E-06	2	(non-genic)	0.299	0.126	0.214
Min_Temp_November	BARC_1.01_Gm_04_6006337_T_C	2.13E-05	1	(non-genic)	0.161	0.347	0.061
Min_Temp_November	BARC_1.01_Gm_13_26065067_G_A	3.43E-05	0	(non-genic)	0.146	0.174	0.152
Min_Temp_October	BARC_1.01_Gm_01_26114693_T_C	6.40E-06	10	(non-genic)	0.01	0	0.216
Min_Temp_October	BARC_1.01_Gm_13_22126286_G_T	2.04E-05	11	(non-genic)	0.26	0.239	0.362
Min_Temp_October	BARC_1.01_Gm_02_5343214_T_C	4.89E-05	0	(non-genic)	0.431	0.484	0.145
Min_Temp_September	BARC_1.01_Gm_01_26114693_T_C	7.63E-06	10	(non-genic)	0.01	0	0.216
Min_Temp_September	BARC_1.01_Gm_04_48092000_T_C	6.95E-05	0	(non-genic)	0.316	0.318	0.458
Min_Temp_September	BARC_1.01_Gm_13_22126286_G_T	7.88E-05	11	(non-genic)	0.26	0.239	0.362

Organic_Matter_Subsoil	BARC_1.01_Gm_17_1729589_G_A	1.70E-04	2	CDS	0.137	0	0.182
Organic_Matter_Subsoil	BARC_1.01_Gm_20_39001833_A_G	3.74E-04	1	(non-genic)	0.125	0.031	0.274
Organic_Matter_Subsoil	BARC_1.01_Gm_04_4137268_A_G	3.92E-04	3	(non-genic)	0.391	0.021	0.125
Organic_Matter_Topsoil	BARC_1.01_Gm_17_1729589_G_A	1.18E-04	2	CDS	0.137	0	0.182
Organic_Matter_Topsoil	BARC_1.01_Gm_20_39001833_A_G	3.68E-04	1	(non-genic)	0.125	0.031	0.274
Organic_Matter_Topsoil	BARC_1.01_Gm_04_4137268_A_G	4.05E-04	3	(non-genic)	0.391	0.021	0.125
Percent_Clay_Subsoil	BARC_1.01_Gm_13_35807751_T_C	9.91E-06	1	(non-genic)	0.203	0	0.338
Percent_Clay_Subsoil	BARC_1.01_Gm_10_36671204_C_A	1.05E-05	1	(non-genic)	0.3	0.462	0.452
Percent_Clay_Subsoil	BARC_1.01_Gm_18_61525330_T_C	2.08E-05	0	(non-genic)	0.292	0.104	0.174
Percent_Clay_Topsoil	BARC_1.01_Gm_10_36671204_C_A	4.98E-06	1	(non-genic)	0.3	0.462	0.452
Percent_Clay_Topsoil	BARC_1.01_Gm_13_35807751_T_C	1.79E-05	1	(non-genic)	0.203	0	0.338
Percent_Clay_Topsoil	BARC_1.01_Gm_20_8269592_A_C	1.92E-05	0	(non-genic)	0.104	0.078	0.13
Percent_Sand_Subsoil	BARC_1.01_Gm_14_23750665_G_A	9.15E-09	4	(non-genic)	0	0	0.061
Percent_Sand_Subsoil	BARC_1.01_Gm_14_29620151_A_C	7.53E-05	0	(non-genic)	0.031	0	0.409
Percent_Sand_Subsoil	BARC_1.01_Gm_12_34354802_G_T	8.09E-05	1	(non-genic)	0.005	0	0.253
Percent_Sand_Topsoil	BARC_1.01_Gm_14_23750665_G_A	1.42E-08	4	(non-genic)	0	0	0.061
Percent_Sand_Topsoil	BARC_1.01_Gm_12_34354802_G_T	6.16E-05	1	(non-genic)	0.005	0	0.253
Percent_Sand_Topsoil	BARC_1.01_Gm_18_18578214_A_G	1.33E-04	0	(non-genic)	0.274	0.266	0.224

Percent_Silt_Subsoil	BARC_1.01_Gm_14_23750665_G_A	5.54E-06	4	(non-genic)	0	0	0.061
Percent_Silt_Subsoil	BARC_1.01_Gm_04_4137268_A_G	2.67E-05	3	(non-genic)	0.391	0.021	0.125
Percent_Silt_Subsoil	BARC_1.01_Gm_01_26114693_T_C	4.83E-05	10	(non-genic)	0.01	0	0.216
Percent_Silt_Topsoil	BARC_1.01_Gm_14_23750665_G_A	2.95E-06	4	(non-genic)	0	0	0.061
Percent_Silt_Topsoil	BARC_1.01_Gm_04_4137268_A_G	1.33E-05	3	(non-genic)	0.391	0.021	0.125
Percent_Silt_Topsoil	BARC_1.01_Gm_01_26114693_T_C	3.05E-05	10	(non-genic)	0.01	0	0.216
pH_Subsoil	BARC_1.01_Gm_04_3461538_T_C	9.16E-06	1	(non-genic)	0.063	0.105	0.231
pH_Subsoil	BARC_1.01_Gm_03_36396038_A_C	3.13E-05	1	(non-genic)	0.442	0.177	0.349
pH_Subsoil	BARC_1.01_Gm_19_48133688_A_C	3.63E-05	0	3UTR	0.163	0.332	0.268
pH_Topsoil	BARC_1.01_Gm_04_3461538_T_C	2.51E-06	1	(non-genic)	0.063	0.105	0.231
pH_Topsoil	BARC_1.01_Gm_03_36396038_A_C	3.10E-05	1	(non-genic)	0.442	0.177	0.349
pH_Topsoil	BARC_1.01_Gm_08_15933908_A_G	3.36E-05	0	Intron	0.323	0.432	0.125
Precipitation_Coldest_Quarter	BARC_1.01_Gm_02_41646843_T_C	2.01E-07	2	(non-genic)	0.042	0.01	0.492
Precipitation_Coldest_Quarter	BARC_1.01_Gm_16_30075422_T_C	2.85E-07	2	Intron	0.311	0.332	0.14
Precipitation_Coldest_Quarter	BARC_1.01_Gm_02_41663747_A_G	2.99E-07	2	(non-genic)	0.042	0.01	0.386
Precipitation_Driest_Month	BARC_1.01_Gm_13_31206278_G_A	1.36E-06	2	CDS	0.096	0.463	0.032
Precipitation_Driest_Month	BARC_1.01_Gm_02_8706603_A_G	6.34E-06	1	(non-genic)	0.401	0.37	0.182
Precipitation_Driest_Month	BARC_1.01_Gm_02_47357235_C_T	1.39E-05	1	Intron	0.378	0.406	0.13

Precipitation_Driest_Quarter	BARC_1.01_Gm_13_31206278_G_A	2.24E-06	2	CDS	0.096	0.463	0.032
Precipitation_Driest_Quarter	BARC_1.01_Gm_02_8706603_A_G	4.96E-06	1	(non-genic)	0.401	0.37	0.182
Precipitation_Driest_Quarter	BARC_1.01_Gm_02_47357235_C_T	6.61E-06	1	Intron	0.378	0.406	0.13
Precipitation_Seasonality	BARC_1.01_Gm_06_1780489_C_T	1.24E-07	0	(non-genic)	0.031	0	0.468
Precipitation_Seasonality	BARC_1.01_Gm_17_12689235_A_G	4.85E-07	1	(non-genic)	0.011	0	0.497
Precipitation_Seasonality	BARC_1.01_Gm_17_1729589_G_A	1.37E-06	2	CDS	0.137	0	0.182
Precipitation_Warmest_Quarter	BARC_1.01_Gm_20_45774939_G_A	4.29E-08	1	(non-genic)	0.223	0.255	0.187
Precipitation_Warmest_Quarter	BARC_1.01_Gm_13_26562075_C_T	2.05E-07	2	(non-genic)	0.286	0.396	0.346
Precipitation_Warmest_Quarter	BARC_1.01_Gm_02_9039246_T_C	6.06E-07	0	(non-genic)	0.137	0.415	0.177
Precipitation_Wettest_Month	BARC_1.01_Gm_13_26562075_C_T	5.69E-09	2	(non-genic)	0.286	0.396	0.346
Precipitation_Wettest_Month	BARC_1.01_Gm_13_26558416_G_T	3.57E-07	0	(non-genic)	0.289	0.396	0.306
Precipitation_Wettest_Month	BARC_1.01_Gm_15_12466753_G_A	1.06E-06	0	(non-genic)	0.285	0.398	0.338
Precipitation_Wettest_Quarter	BARC_1.01_Gm_13_26562075_C_T	2.06E-08	2	(non-genic)	0.286	0.396	0.346
Precipitation_Wettest_Quarter	BARC_1.01_Gm_20_45774939_G_A	4.86E-08	1	(non-genic)	0.223	0.255	0.187
Precipitation_Wettest_Quarter	BARC_1.01_Gm_08_2254106_G_A	1.42E-07	1	CDS	0.224	0.01	0.332
Temperature_Annual_Range	BARC_1.01_Gm_05_4252974_A_G	2.05E-06	1	(non-genic)	0	0	0.095
Temperature_Annual_Range	BARC_1.01_Gm_17_12689235_A_G	3.70E-06	1	(non-genic)	0.011	0	0.497
Temperature_Annual_Range	BARC_1.01_Gm_11_21381584_T_C	1.43E-05	0	(non-genic)	0.326	0.01	0.28

Temperature_Seasonality	BARC_1.01_Gm_08_42562917_A_G	2.44E-06	0	(non-genic)	0.063	0	0
Temperature_Seasonality	BARC_1.01_Gm_04_4217329_G_T	1.31E-05	0	(non-genic)	0.474	0.021	0.176
Temperature_Seasonality	BARC_1.01_Gm_08_14197379_T_C	2.84E-05	0	(non-genic)	0.229	0.011	0.118

Table S3. Significant markers from SPA analysis.

SNP Name	Chr.	Location	Selection score	Near Gene	Arabidopsis Top hit	GO Molecular Function	GO Biological Process
BARC_1.01_Gm02_42655797_T_C	2	39,585,022	6.64	Glyma.02G210600	AT1G72200	zinc ion binding	cellular response to iron ion starvation
BARC_1.01_Gm03_43808836_T_C	3	41,805,260	6.59	Glyma.03G211900	AT1G02970	protein binding	DNA endoreduplication
BARC_1.01_Gm04_5572742_G_A	4	5,639,340	6.81	Glyma.04G067500	AT3G47650	-	heat shock protein binding, unfolded protein binding
BARC_1.01_Gm04_12410533_A_G	4	13,499,694	6.55	Glyma.04G116500	AT2G17390	-	regulation of transcription, DNA-templated
BARC_1.01_Gm04_37448392_G_A	4	40,597,994	7.07	Glyma.04G163600	AT5G07940		
BARC_1.01_Gm04_37474521_G_T	4	40,624,123	7.05				
BARC_1.01_Gm04_37799809_T_C	4	40,944,495	6.55	-			
BARC_1.01_Gm04_37867591_T_G	4	41,009,165	6.63	Glyma.04G164500	AT5G07830	beta-glucuronidase activity	unidimensional cell growth
BARC_1.01_Gm04_39001105_C_A	4	42,162,859	6.67	Glyma.04G168100	AT2G23420	nicotinate phosphoribosyltransferase activity	NAD biosynthetic process, nicotinate nucleotide salvage
BARC_1.01_Gm08_14711307_C_A	8	14,641,696	6.83	Glyma.08G182500	AT1G25270	transmembrane transporter activity	
BARC_1.01_Gm08_23555342_T_C	8	23,473,832	6.73	Glyma.08G260000	-		
BARC_1.01_Gm08_23691942_G_A	8	23,614,953	6.66	-			
BARC_1.01_Gm08_33021069_A_G	8	33,629,938	6.64	Glyma.08G267500	-		

BARC_1.01_Gm08_33515066_A_G	8	34,120,361	6.95	Glyma.08G267800	AT4G20170	transferase activity	cell wall biogenesis
BARC_1.01_Gm11_4966217_C_A	11	4,975,767	7.23	Glyma.11G066300	AT4G36670	carbohydrate transmembrane transporter activity	cation transmembrane transport, glucose import
BARC_1.01_Gm11_7753133_T_C	11	7,763,295	6.94	Glyma.11G102100	AT1G50200	alanine-tRNA ligase activity	alanyl-tRNA aminoacylation
BARC_1.01_Gm11_7753765_T_C	11	7,763,927	6.90	Glyma.11G102100	AT1G50200	alanine-tRNA ligase activity	alanyl-tRNA aminoacylation
BARC_1.01_Gm11_7843684_C_T	11	7,854,392	6.57	Glyma.11G103300	AT3G19830	molecular_function unknown	biological_process unknown
BARC_1.01_Gm13_30479725_C_T	13	31,691,971	6.69	Glyma.13G203000	AT3G08040	antiporter activity	cellular iron ion homeostasis
BARC_1.01_Gm14_5164997_A_G	14	5,275,680	6.79	Glyma.14G064400	AT3G46850	serine-type endopeptidase activity	metabolic process, proteolysis
BARC_1.01_Gm15_10382285_T_C	15	9,417,700	7.39	Glyma.15G119500, Glyma.15G119600, Glyma.15G119700, Glyma.15G119800	AT5G46890.AT5G46900	lipid binding	lipid transport
BARC_1.01_Gm15_10376148_G_A	15	9,423,838	7.21	"	“	“	“

Table S4. Markers found to be significant F_{ST} outliers.

SNP Name	Chr	Location	Near Gene within half life of LD in *G. soja*	Arabidopsis Top hit	Annotated Function
BARC_1.01_Gm04_22830013_G_A	4	28208509	NA	NA	NA
BARC_1.01_Gm04_21921238_A_C	4	29130389	NA	NA	NA
BARC_1.01_Gm08_30225439_G_A	8	30877366	Glyma08g33580	AT4G00720	ATP binding, kinase activity, protein serine/threonine kinase activity
BARC_1.01_Gm08_32890861_G_T	8	33495860	Glyma.08g267200	AT3G51700.1	DNA repair, telomere maintenance
BARC_1.01_Gm09_38799984_C_T	9	41477881	Glyma.09g190200	AT3G29300.1	unknown protein;
BARC_1.01_Gm09_38670296_A_C	9	41346640	Glyma.09g188700	AT3G51895.1	Encodes a chloroplast-localized sulfate transporter.
BARC_1.01_Gm09_38806410_G_A	9	41484307	Glyma.09g190400	AT4G32350.1	Regulator of Vps4 activity in the MVB pathway protein
BARC_1.01_Gm09_41854884_C_T	9	45050918	Glyma.09g225600	AT3G51950.1	Nucleotide-binding, alpha-beta plait:IPR012677(1)
BARC_1.01_Gm09_41853283_T_C	9	45049317	Glyma.09g225600	AT3G51950.1	Nucleotide-binding, alpha-beta plait:IPR012677(1)
BARC_1.01_Gm10_38702370_C_A	10	39250548	Glyma.10g158600	AT3G05010.1	unknown protein;
BARC_1.01_Gm11_10249349_G_A	11	10280660	Glyma.11g134700, Glyma.11g134800	AT2G26975.1	Ctr copper transporter family; FUNCTIONS IN: copper ion transmembrane transporter activity;
BARC_1.01_Gm12_36126563_G_A	12	36117254	Glyma.12g199800	AT1G03940.1	HXXXD-type acyl-transferase family protein; FUNCTIONS IN: transferase activity, transferring acyl groups other than amino-acyl groups, transferase activity;
BARC_1.01_Gm13_42200703_A_C	13	43640741	Glyma.13g345800	AT2G43465.1	RNA-binding ASCH domain protein;

BARC_1.01_Gm13_2147651_G_A	13	19719551	Glyma.13g085600	NA	NA
BARC_1.01_Gm13_24666414_C_A	13	26152104	Glyma.13g148000	AT5G03560.1	Tetratricopeptide repeat (TPR)-like superfamily protein
BARC_1.01_Gm13_36928707_G_A	13	38069908	Glyma.13g279500	AT4G27435.1	unknown protein;
BARC_1.01_Gm14_32429662_A_C	14	15579394	Glyma.14g117800	AT4G27680.1	P-loop containing nucleoside triphosphate hydrolases superfamily protein
BARC_1.01_Gm15_10382285_T_C	15	9417700	Glyma.15G119500, Glyma.15G119600, Glyma.15G119700, Glyma.15G119800	AT5G46890.AT5G46900	lipid binding/transport
BARC_1.01_Gm15_10376148_G_A	15	9423838	Glyma.15G119500, Glyma.15G119600, Glyma.15G119700, Glyma.15G119800	AT5G46890.AT5G46900	lipid binding/transport
BARC_1.01_Gm15_8042864_C_T	15	8074977	Glyma.15g103400, Glyma.15g103500, Glyma.15g103300	AT3G63088.1 , AT3G06170.1	ROTUNDIFOLIA like 14, Serinc-domain containing serine and sphingolipid biosynthesis protein
BARC_1.01_Gm15_9365338_C_T	15	9509216	Glyma.15g120200	AT2G24960.2	unknown protein;
BARC_1.01_Gm20_37503728_A_C	20	38608664	Glyma.20g147600	AT3G22490.1	Seed maturation protein;

Table S5. Genomic location enrichment analysis, bold text indicates a significant enrichment.

	Number across all SNPs	Percent of SNPs	99% CI	Significant in Environmental Association	Percent of significant associations	SPA outlier	Percent of SPA outliers	F_{ST} outlier	Percent of F_{ST} outlier
Genic	9644	29.75%	26.4-33.0%	26	**23.64%**	9	**42.86%**	3	13.60%
3UTR	436	1.40%	0.01-2.3%	4	**3.64%**	0	0.00%	0	0.00%
5UTR	465	1.40%	0.01-2.3%	0	0.00%	0	0.00%	0	0.00%
CDS	4192	12.90%	10.6-15.3%	11	**10.00%**	3	13.60%	3	13.60%
Intron	4551	14.00%	11.6-16.9%	11	**10.00%**	6	**27.20%**	0	0.00%
Non-genic	22772	70.2	66.3-73.8%	84	**76.36%**	13	**59.10%**	19	**86.60%**
All	32416	NA	NA	110 (unique)	NA	22	NA	22	NA

SUPPLEMENTARY FIGURES

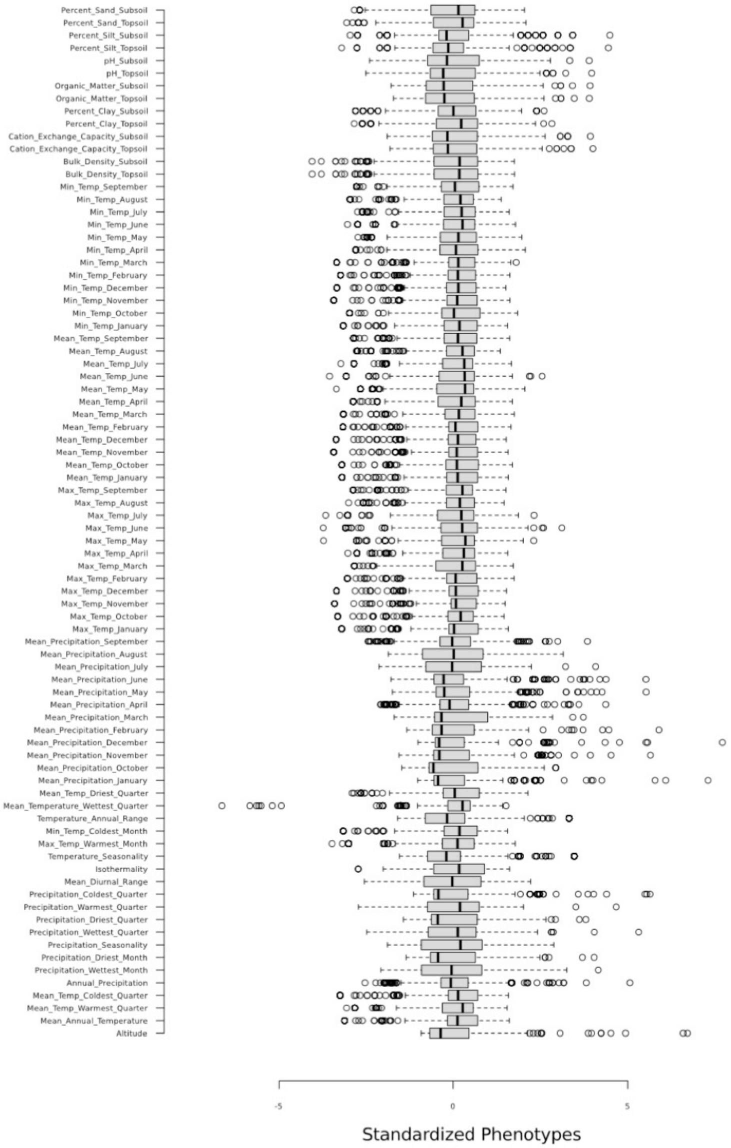


Figure S1. Standardized distributions of biophysical (soil) and bioclimatic variables. Raw values found in table S3.

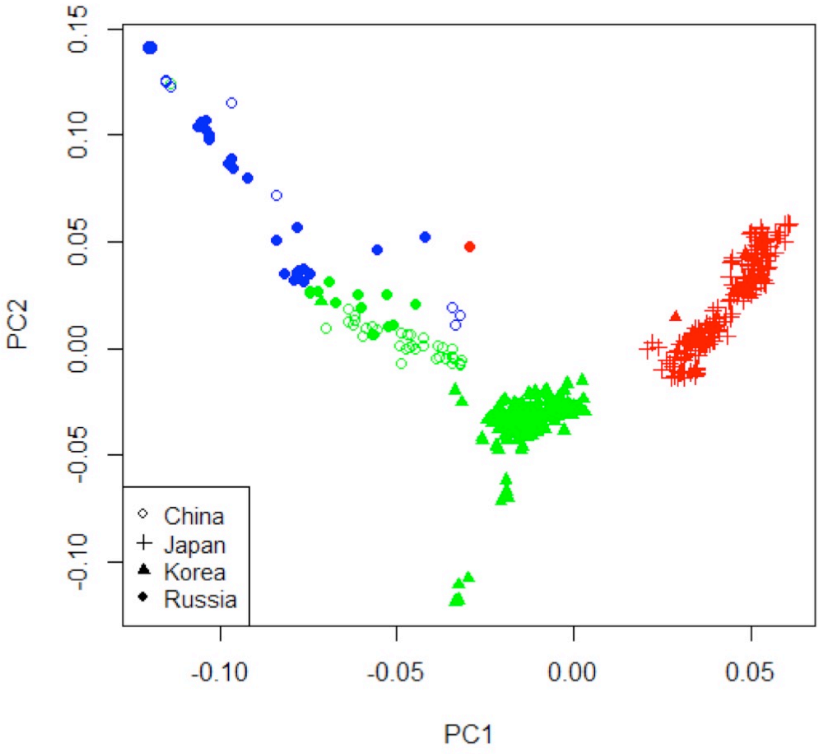


Figure S2. PC1 and PC2 of 533 individuals among *G. soja* accessions. Samples are colored by structure assignment (Blue is Mainland North, Green is Mainland South, and Red is Island) with shapes for each country of origin. The first PC explained 5.1% of the variation and separated samples in an east to west gradient. The second PC explained 2.7% of the variation and corresponded more generally to a north south separation.

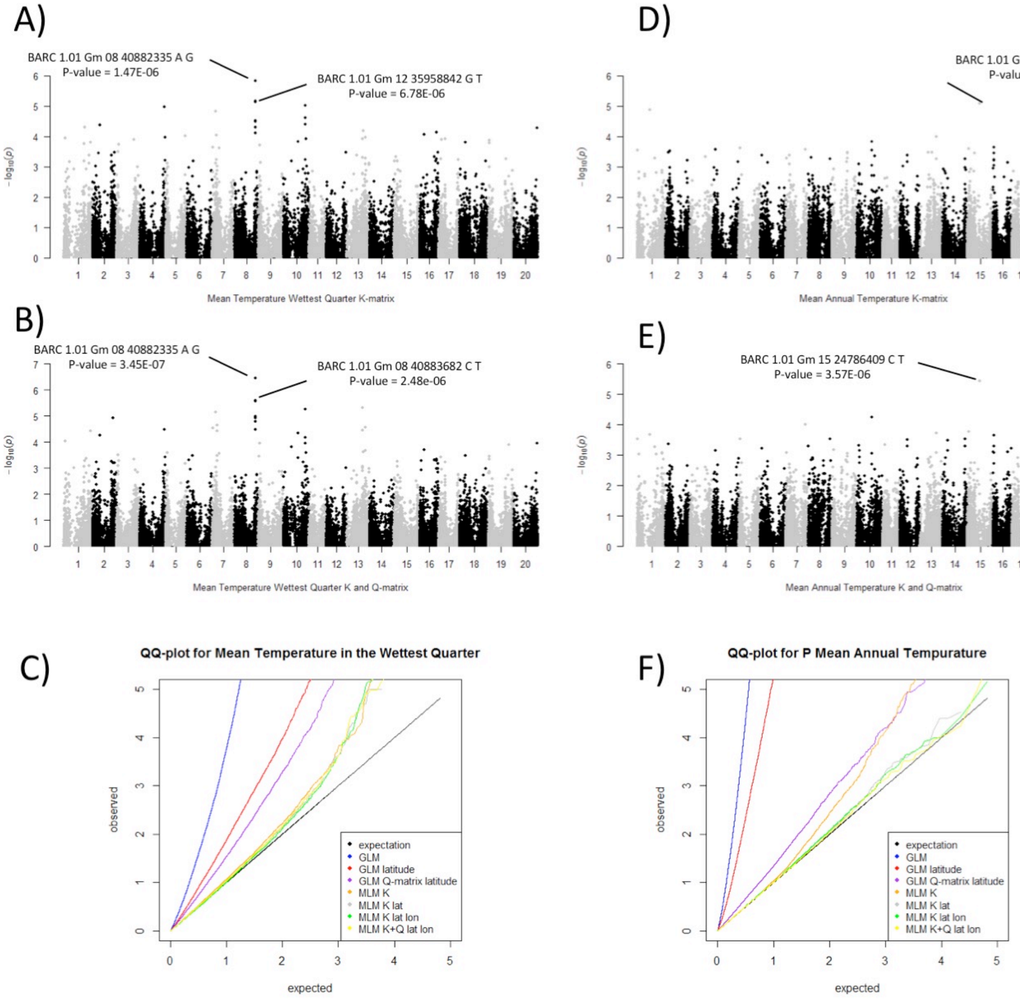


Figure S3. Mixed-model association mapping results for Mean Temperature Wettest Quarter. A) Manhattan plot of the association results when controlled for K-matrix (relatedness) and latitude for Mean Temperature in the Wettest Quarter. B) Manhattan plot of the association results when controlled for K-matrix, Q-matrix (population structure), and latitude for Mean Temperature in the Wettest Quarter. The top two markers were the exact same using both methods. The significance of the markers changed in the models. C) Quantile-Quantile plot for Mean Temperature in the Wettest Quarter. Here the K + latitude, K + latitude + longitude, K + Q + latitude and Q + K+ latitude + longitude performed similarly, however, the genomic inflation parameter Λ was 1.04 for the K + latitude model and 1.03 for Q + K + latitude model, indicating that the both models performed closely to the expectation of 1. D) Manhattan plot of the association results when controlled for K-matrix (relatedness) and latitude and Mean Annual Temperature. E) Manhattan plot of the association results when controlled for K-matrix, Q-matrix (population structure), and latitude for Mean Annual Temperature. The most significant maker was the same using both methods. The significance of the markers changed in the models. F) Quantile-Quantile plot for Mean Annual Temperature. Here the K + latitude, K + latitude + longitude, K + Q + latitude and Q + K + latitude + longitude, the genomic inflation parameter Λ was 1.04 for the K + latitude model and 1.04 for Q +K + latitude model, indicating that the both models performed closely to the expectation of 1. Across other bioclimatic and biophysical variables the K model had a smaller variance in Λ compared to the Q + K model.

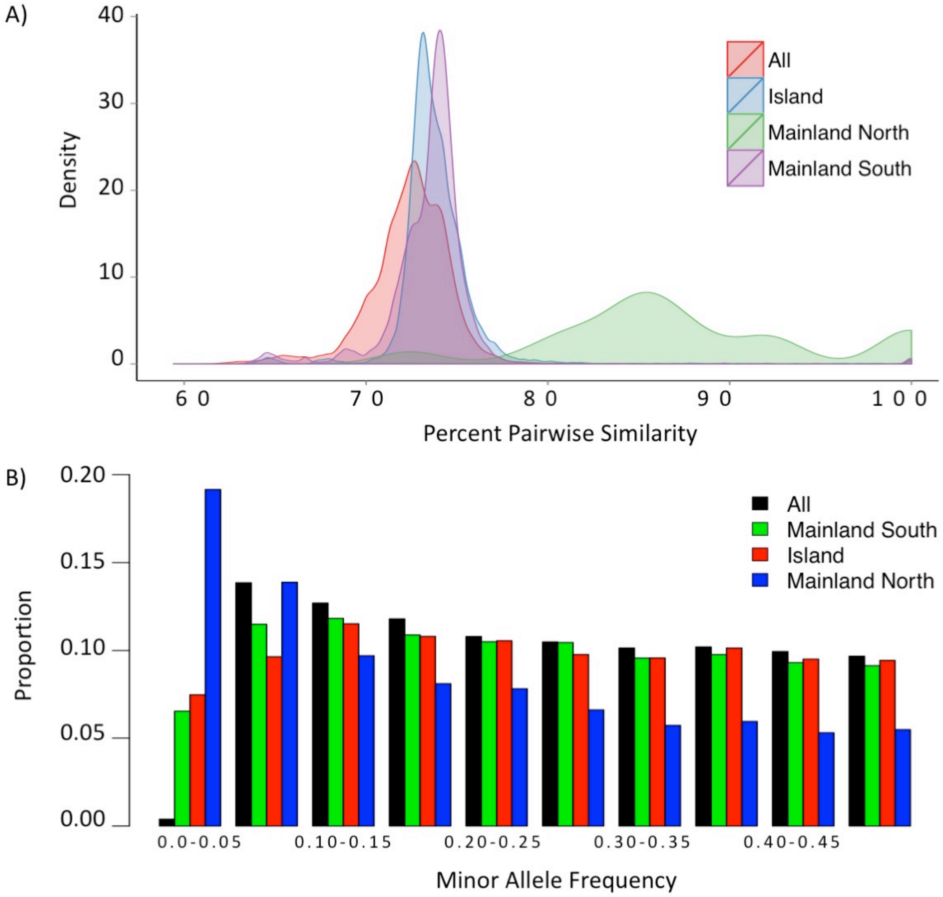


Figure S4: A) Density plots of pairwise similarity for each fastSTRUCTURE cluster of *G. soja* B) Folded site frequency spectrum of all markers in all individuals used in this study.

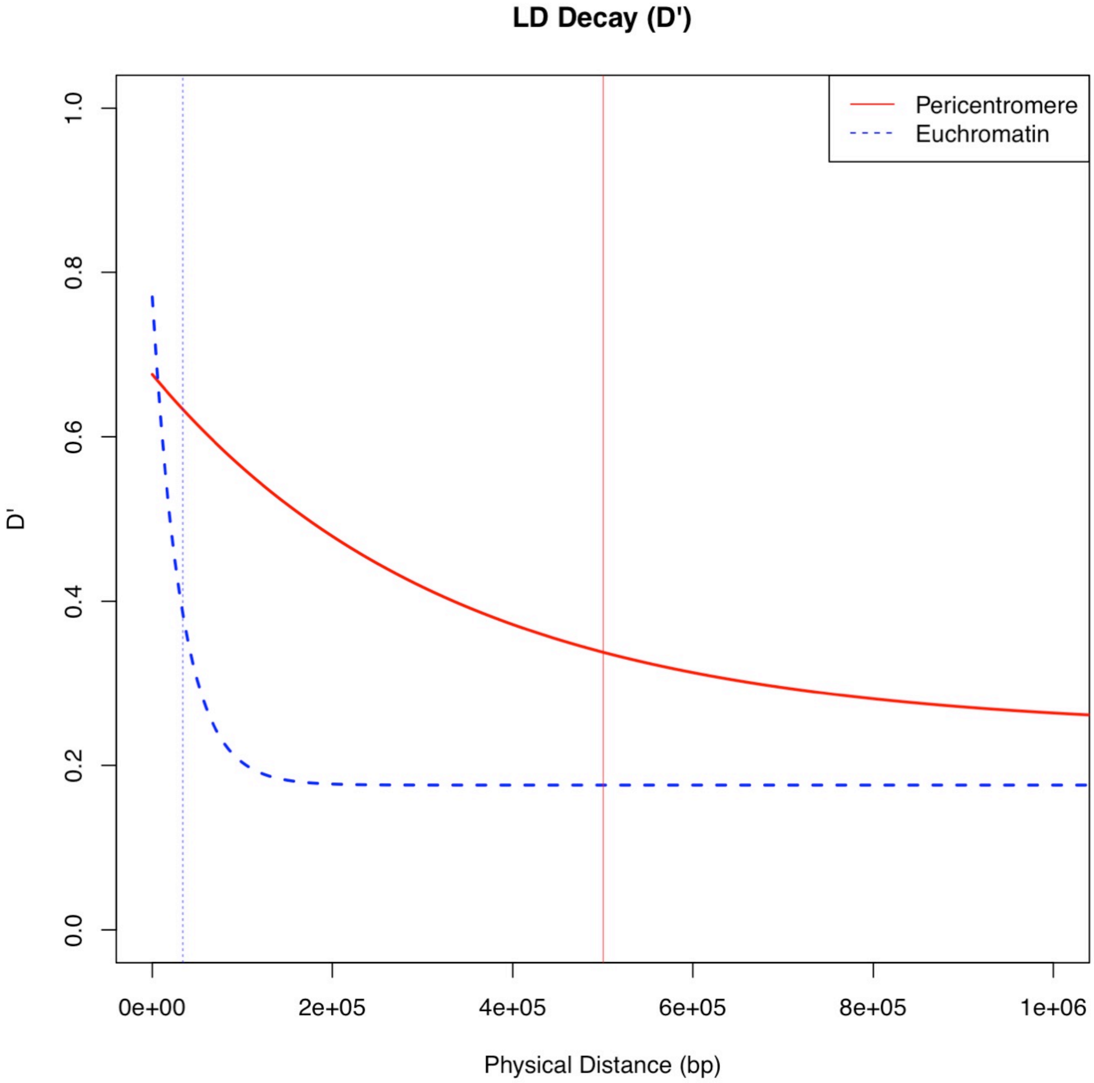


Figure S5. LD decay in G. soja. A) LD decay according to D′ plotted for pericentromere and euchromatin.

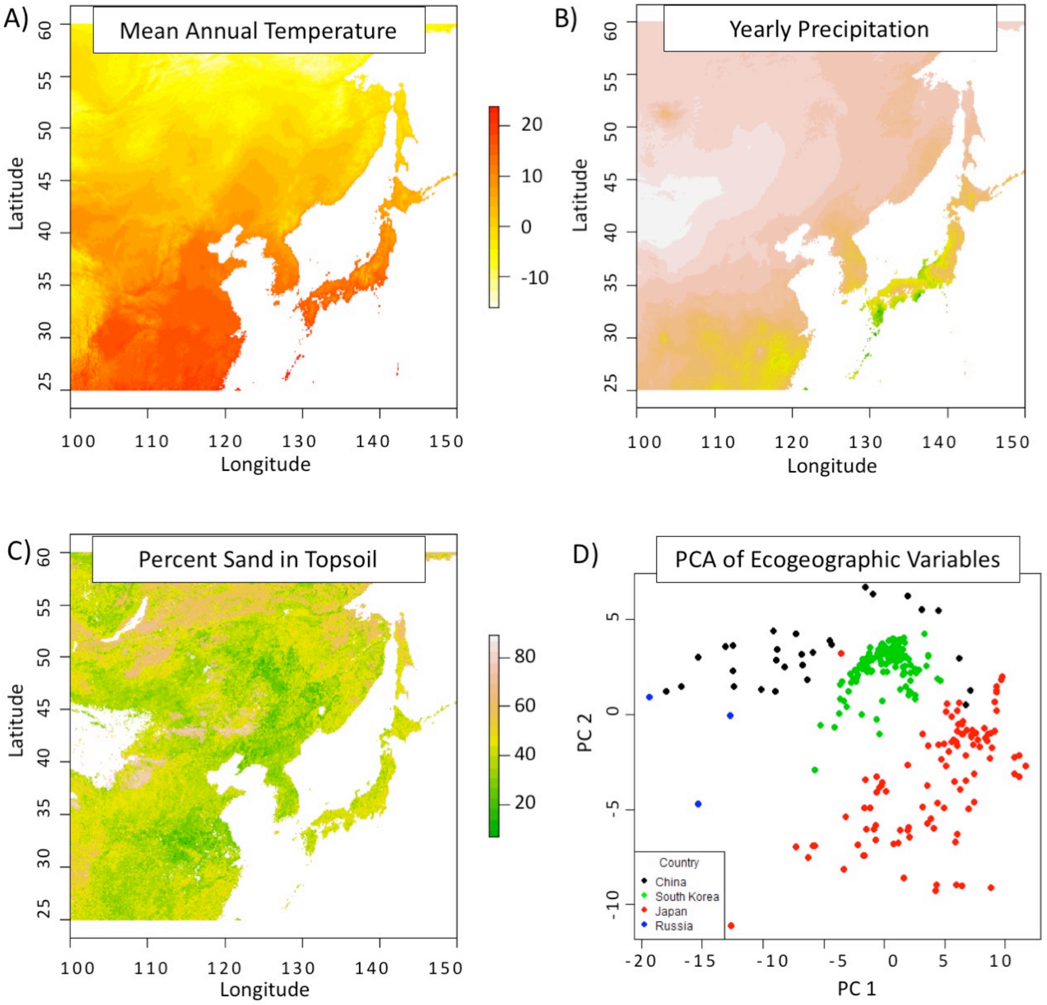


Figure S6. Variation in ecogeographic variables across the range of *G. soja*. A) Mean annual temperature measured in degrees Celsius. B) Yearly precipitation measured in cm. C) Percent sand in top 30 centimeters of soil.(White is no data) D) PC1 and PC2 showing differentiation of climate and soil conditions at the sampling locations of 533 *G. soja* accessions. Dots are colored by country. The first four principle components (PCs) explained 86.3% of the bioclimatic and biophysical variation across the range of *G. soja* with much of the variation being related to temperature and precipitation seasonality.

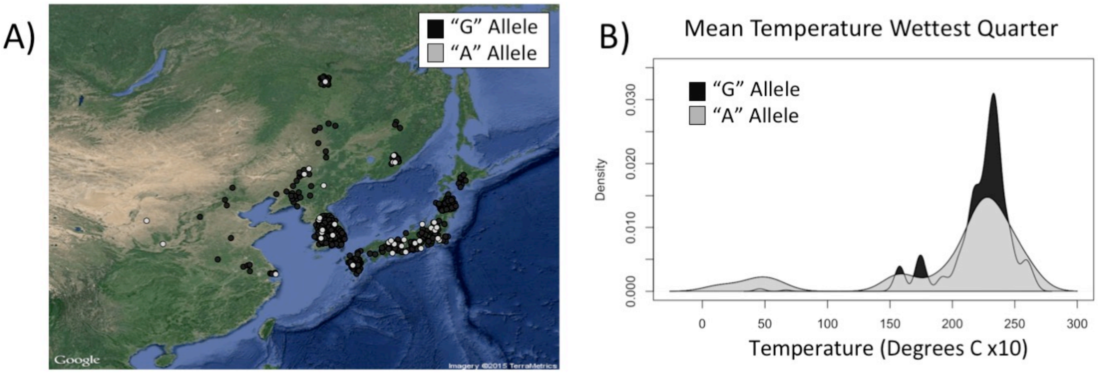
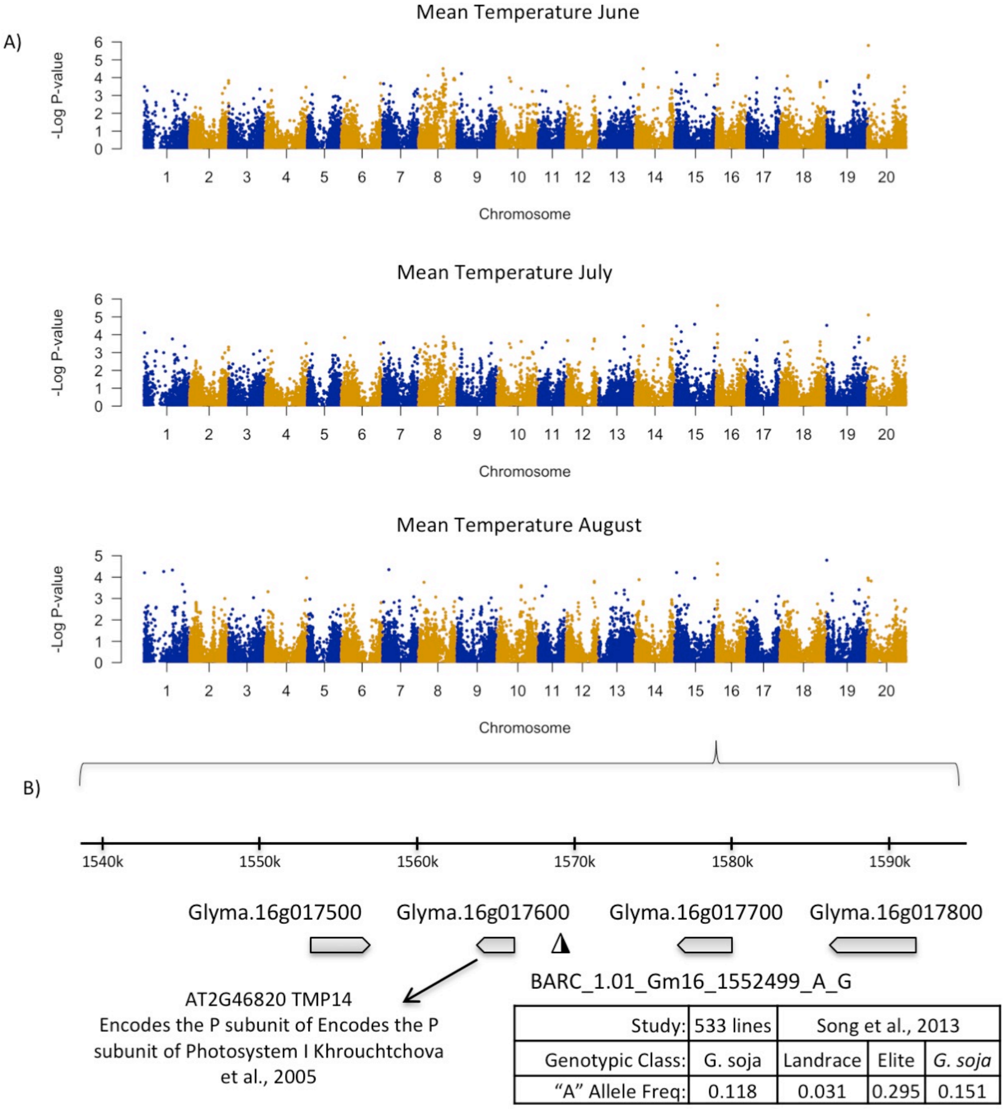


Figure S7. Investigation of SNP: BARC_1.01_Gm08_40882335_A_G distribution, the most significant marker associated with Mean Temperature Wettest Quarter. A) Geographic location of individuals with the reference allele "A" (light gray) or non-reference allele "G" (dark gray) with jitter added to show overlapping samples. Individuals with missing genotyping data are not shown. B) Density plot of allele frequency distribution for Mean Temperature Wettest Quarter. The reference allele "A" individuals are shaded in light gray overlaid with the non-reference allele "G" individuals in dark gray. The average Mean Temperature Wettest Quarter is 20.04 C for the 34 individuals carrying the "A" allele and 22.22 C for the 496 individuals carrying the "G" allele.



Study:	533 lines	Song et al., 2013		
Genotypic Class:	G. soja	Landrace	Elite	*G. soja*
"A" Allele Freq:	0.118	0.031	0.295	0.151

Figure S8. One marker, BARC_1.01_Gm16_1552499_A_G, was found significant in 11 temperature related bioclimatic variables. This included Max Temperature Warmest Month, Mean Temperature Warmest Quarter, Maximum Temperature June, July, and August, Mean Temperature May, June, July, and August, and Minimum Temperature June and July. A) Manhattan plots of genome-wide association results for Mean June, July, and August Temperature. B) Zoom in on 60 kb region around the significant marker BARC 1.01 Gm16 1552499 A G. The Arabidopsis top hit for the nearest gene, Glyma.16g017600, is AT2G46820 TMP14 , which encodes the P subunit of Photosystem I (Khrouchtchova et al., 2005). The "A" allele was only found to be rare in landraces based on a previous study.

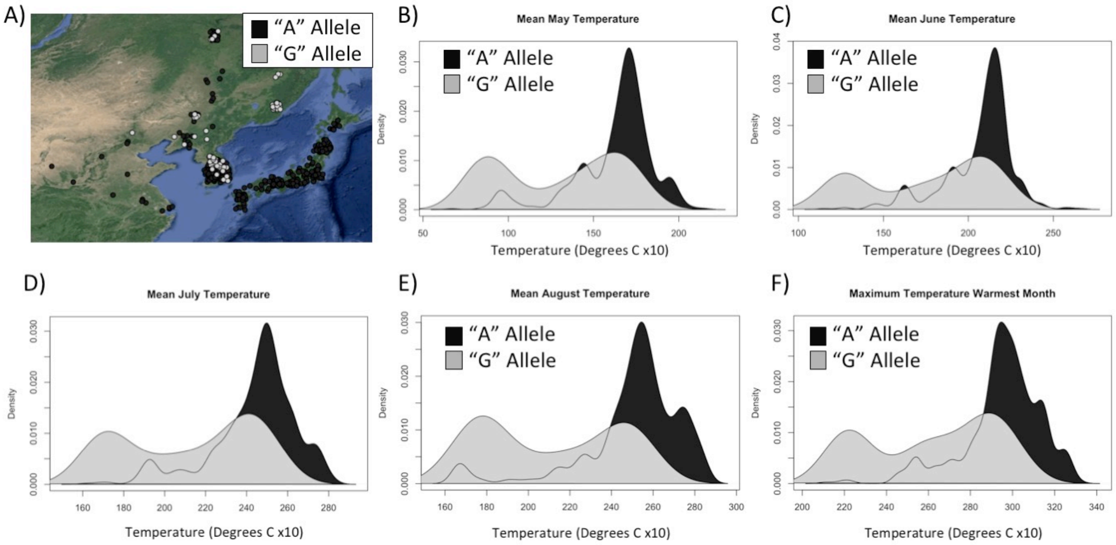
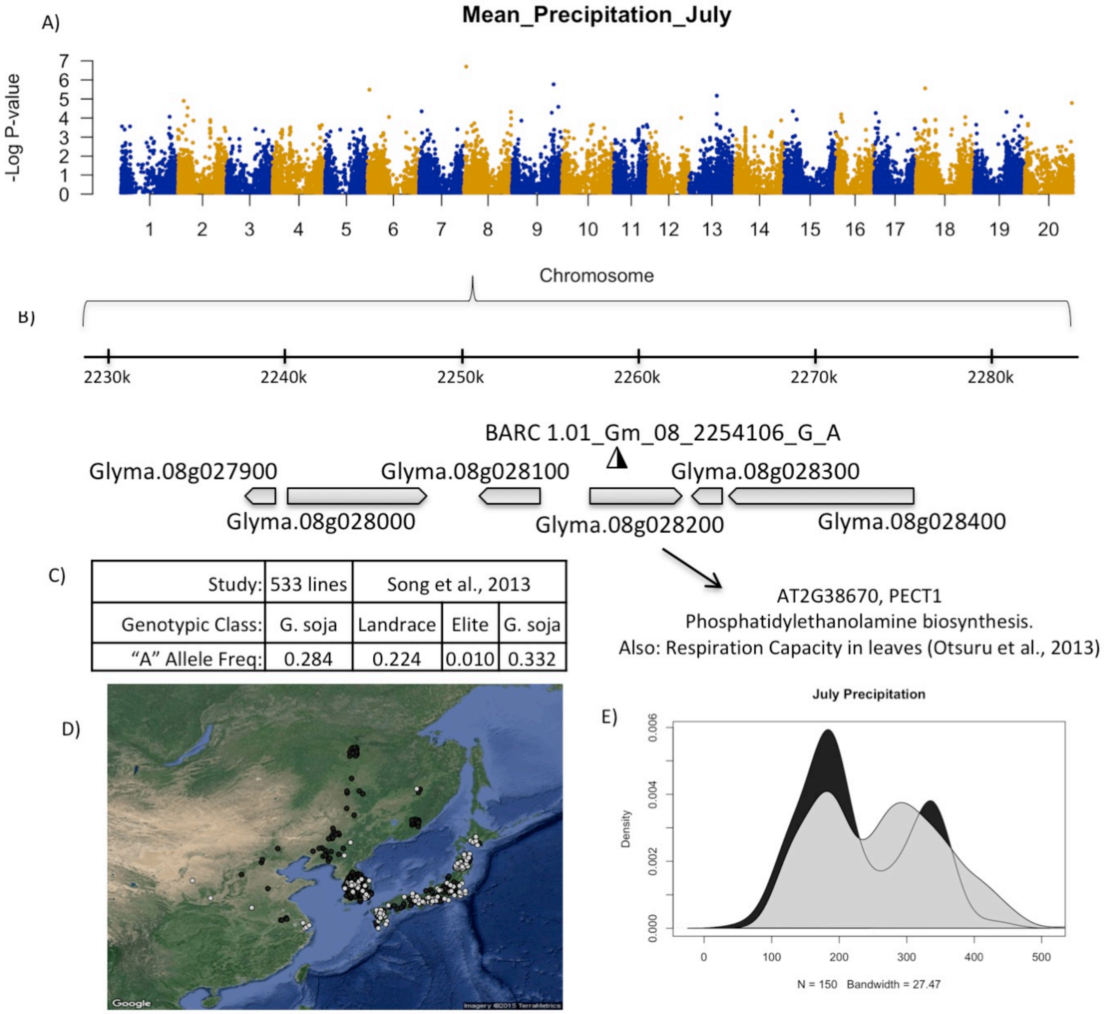


Figure S9. Investigation of SNP: BARC_1.01_Gm16_1552499_A_G distribution. This marker was signficant for Max Temperature Warmest Month, Mean Temperature Warmest Quarter, Maximum Temperature June, July, and August, Mean Temperature May, June, July, and August, and Minimum Temperature June and July. A) Geographic location of individuals with the reference allele "A" (Dark gray) or non-reference allele "G" (light gray) with jitter added to show overlapping samples. Individuals with missing genotyping data are not shown. B)-G) Density plots of some of the environmental variables found in significant association with this SNP. The reference allele "A" individuals are shaded in dark gray overlaid with the non-reference allele "G" individuals in light gray.



Study:	533 lines	Song et al., 2013		
Genotypic Class:	G. soja	Landrace	Elite	G. soja
"A" Allele Freq:	0.284	0.224	0.010	0.332

Figure S10. Genome-wide association significant marker for July Precipitation and Precipitation Wettest Quarter. A) Manhattan plot of July Precipitation association results. B) Zoom in on 60 kb region around the significant marker BARC_1.01_Gm_08_2254106_G_A. The Arabidopsis homolog for the nearest gene, Glyma.08g028200, is AT2G38670, PECT1, involved in Phosphatidylethanolamine biosynthesis but also implicated in respiration capacity in leaves (Otsuru et al., 2013). C) The "A" allele is common in *G. soja* and landraces, but rare in elite lines (Song et al., 2013). D) Geographic location of individuals with the allele "A" (light gray) or reference allele "G" (dark gray) with jitter added to show overlapping samples. Individuals with missing genotyping data are not shown. B) Density plot of allele frequency distribution for July Precipitation. The reference allele "G" individuals are shaded in dark gray overlaid with the non-reference allele "A" individuals in dark gray.

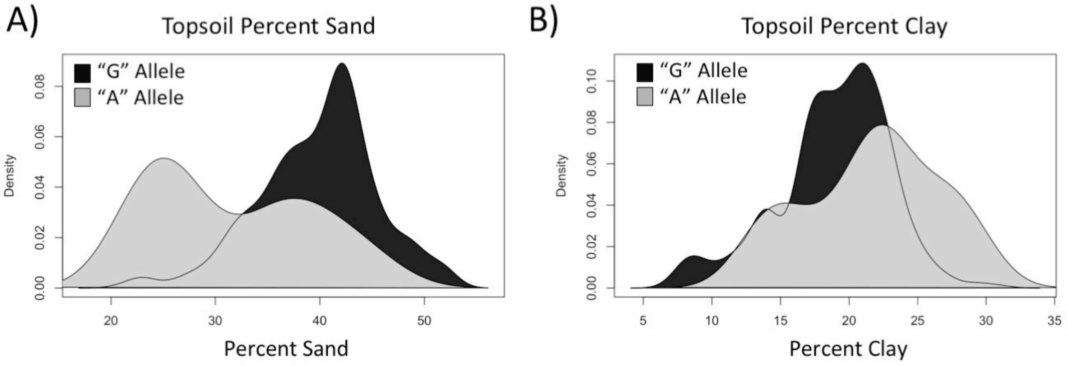
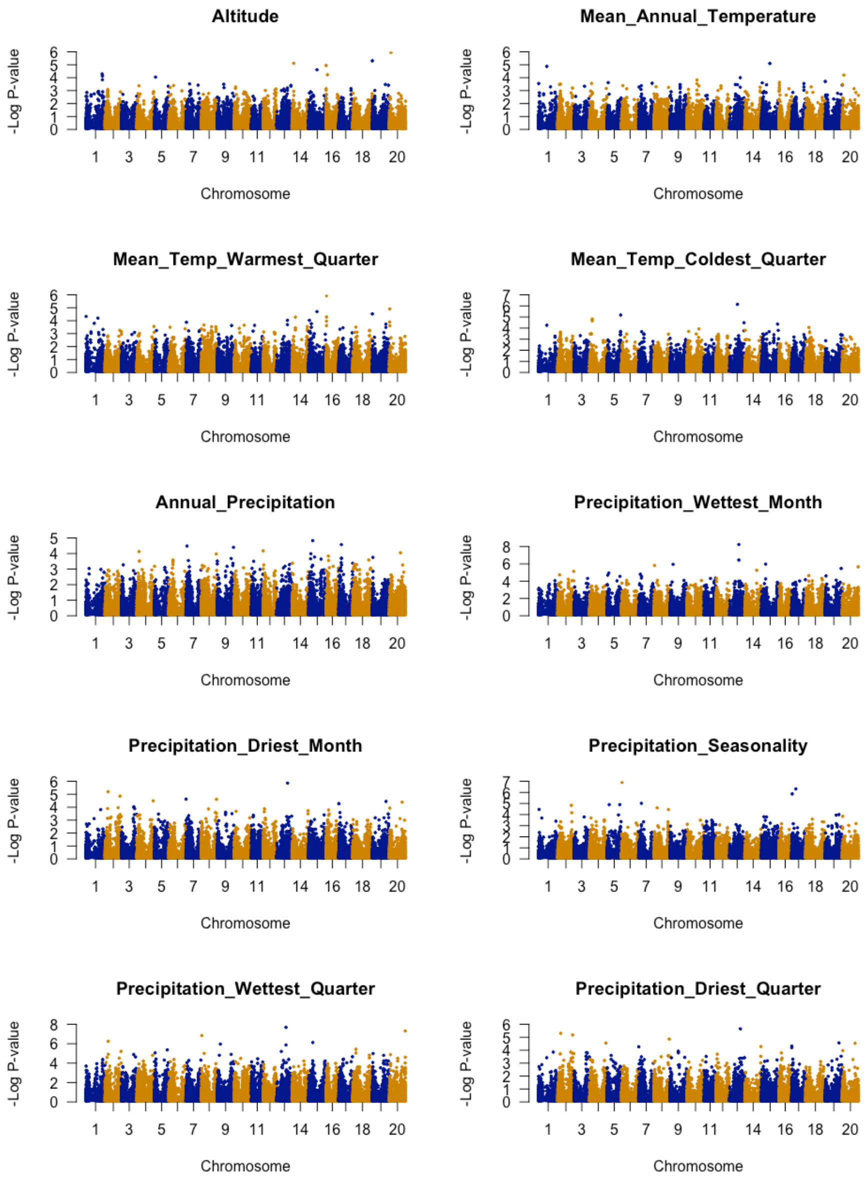
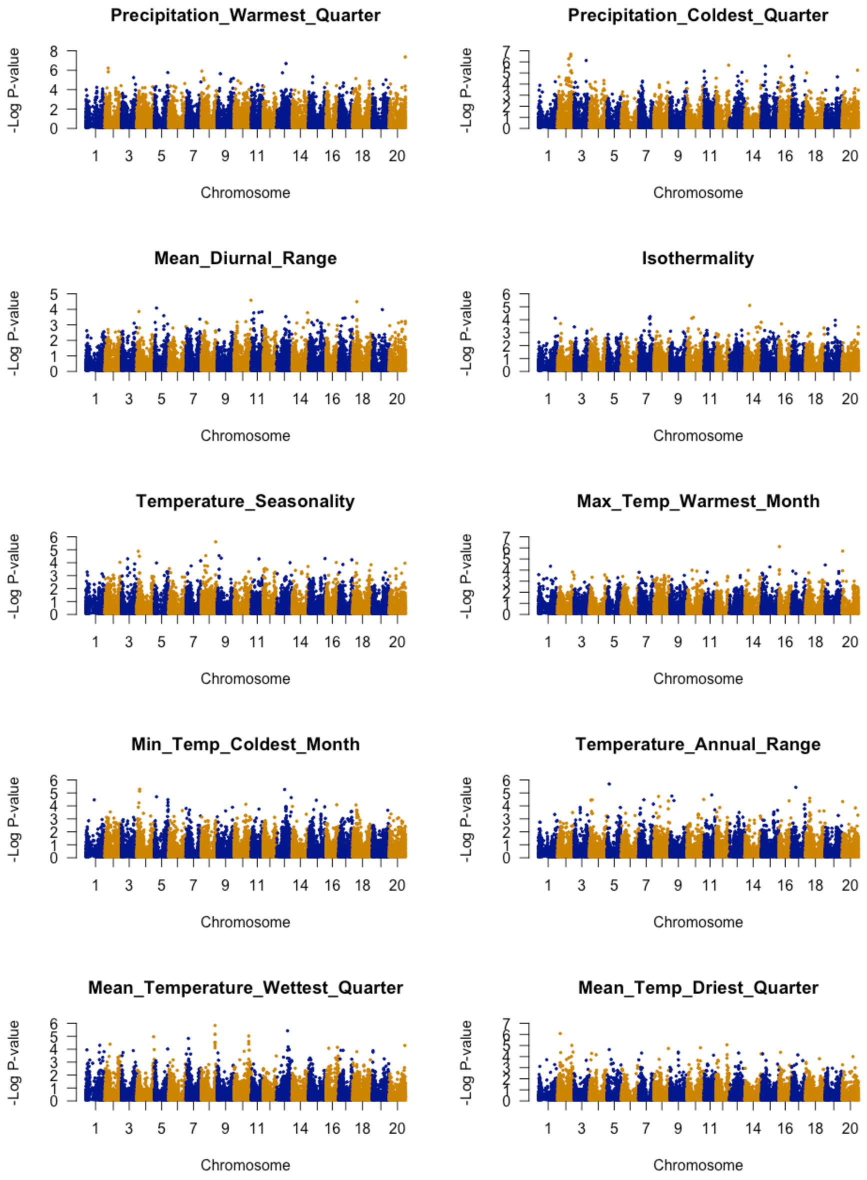
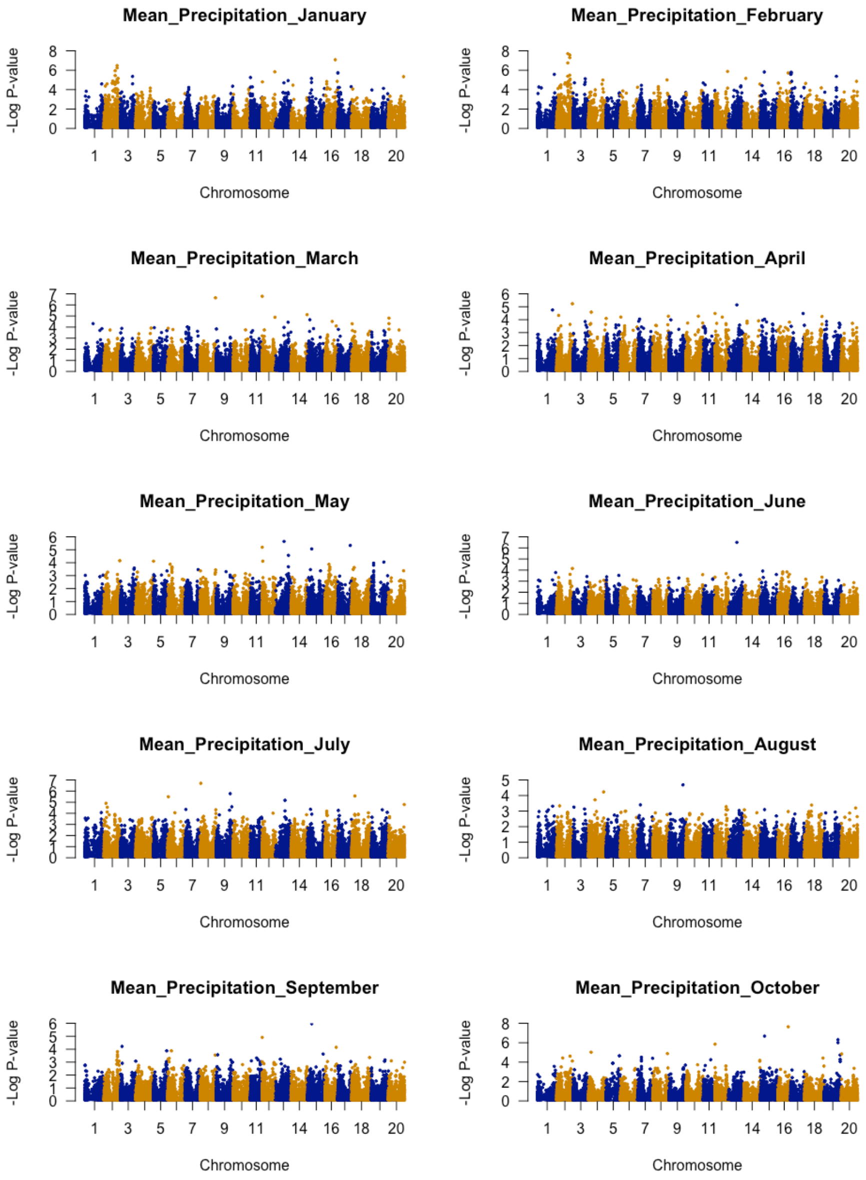


Figure S11. Investigation of SNP: BARC_1.01_Gm14_23750665_G_A distribution. This marker was significant for Percent Sand and Percent Silt in both topsoil and subsoil and Cation Exchange Capacity Topsoil. A) Density plot of raw data for Topsoil Percent Sand. B) Density plot of raw data for Topsoil Percent Clay. The reference allele "G" individuals are shaded in dark gray overlaid with the non-reference allele "A" individuals in light gray.

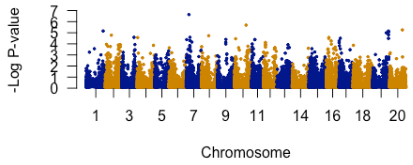
Figure S12. SoySNP50K Bioclimatic and Biophysical association results displayed in Manhattan plots.



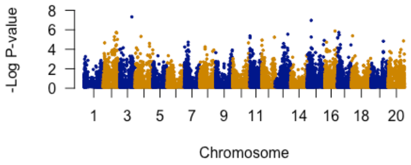

Precipitation_Warmest_Quarter
-Log P-value
0 2 4 6 8
1 3 5 7 9 11 14 16 18 20
Chromosome
Precipitation_Coldest_Quarter
-Log P-value
0 1 2 3 4 5 6 7
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Diurnal_Range
-Log P-value
0 1 2 3 4 5
1 3 5 7 9 11 14 16 18 20
Chromosome
Isothermality
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Temperature_Seasonality
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Max_Temp_Warmest_Month
-Log P-value
0 1 2 3 4 5 6 7
1 3 5 7 9 11 14 16 18 20
Chromosome
Min_Temp_Coldest_Month
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Temperature_Annual_Range
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Temperature_Wettest_Quarter
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Temp_Driest_Quarter
-Log P-value
0 1 2 3 4 5 6 7
1 3 5 7 9 11 14 16 18 20
Chromosome


Mean_Precipitation_January
-Log P-value
0 2 4 6 8
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_February
-Log P-value
0 2 4 6 8
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_March
-Log P-value
0 1 2 3 4 5 6 7
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_April
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_May
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_June
-Log P-value
0 1 2 3 4 5 6 7
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_July
-Log P-value
0 1 2 3 4 5 6 7
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_August
-Log P-value
0 1 2 3 4 5
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_September
-Log P-value
0 1 2 3 4 5 6
1 3 5 7 9 11 14 16 18 20
Chromosome
Mean_Precipitation_October
-Log P-value
0 2 4 6 8
1 3 5 7 9 11 14 16 18 20
Chromosome

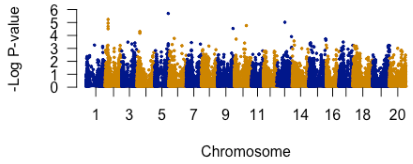
Mean_Precipitation_November



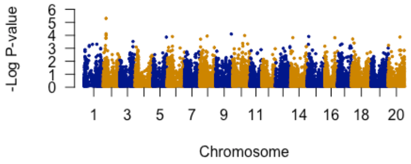
Mean_Precipitation_December



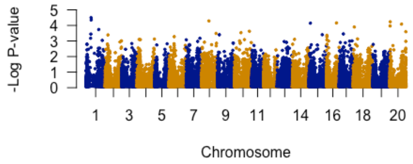
Max_Temp_January



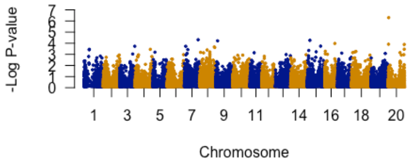
Max_Temp_February



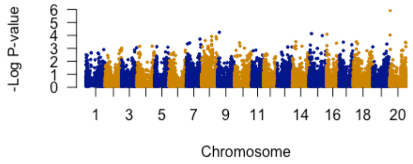
Max_Temp_March



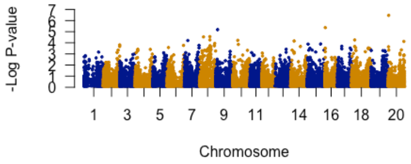
Max_Temp_April



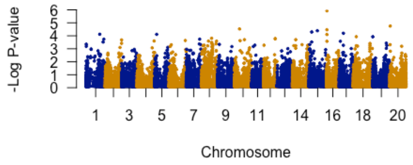
Max_Temp_May



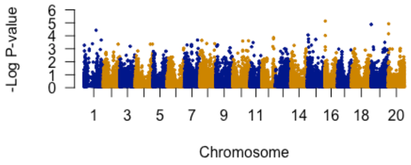
Max_Temp_June



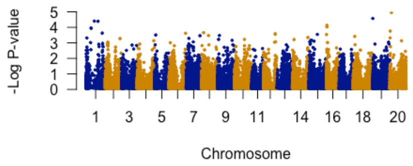
Max_Temp_July



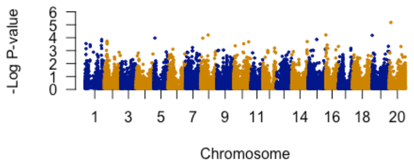
Max_Temp_August



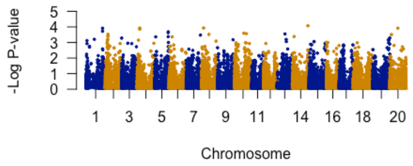
Max_Temp_September



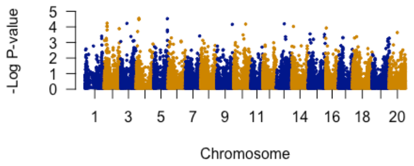
Max_Temp_October



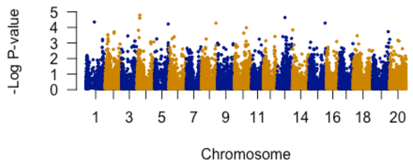
Max_Temp_November



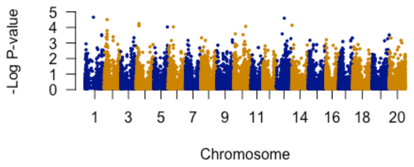
Max_Temp_December



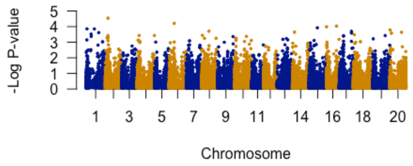
Mean_Temp_January



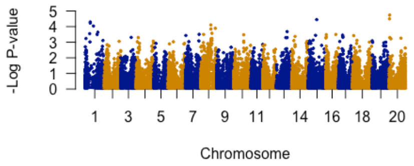
Mean_Temp_February



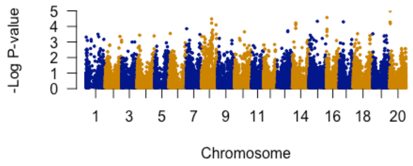
Mean_Temp_March



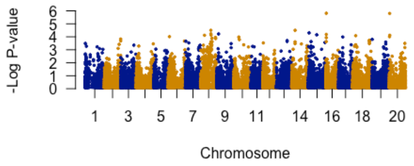
Mean_Temp_April

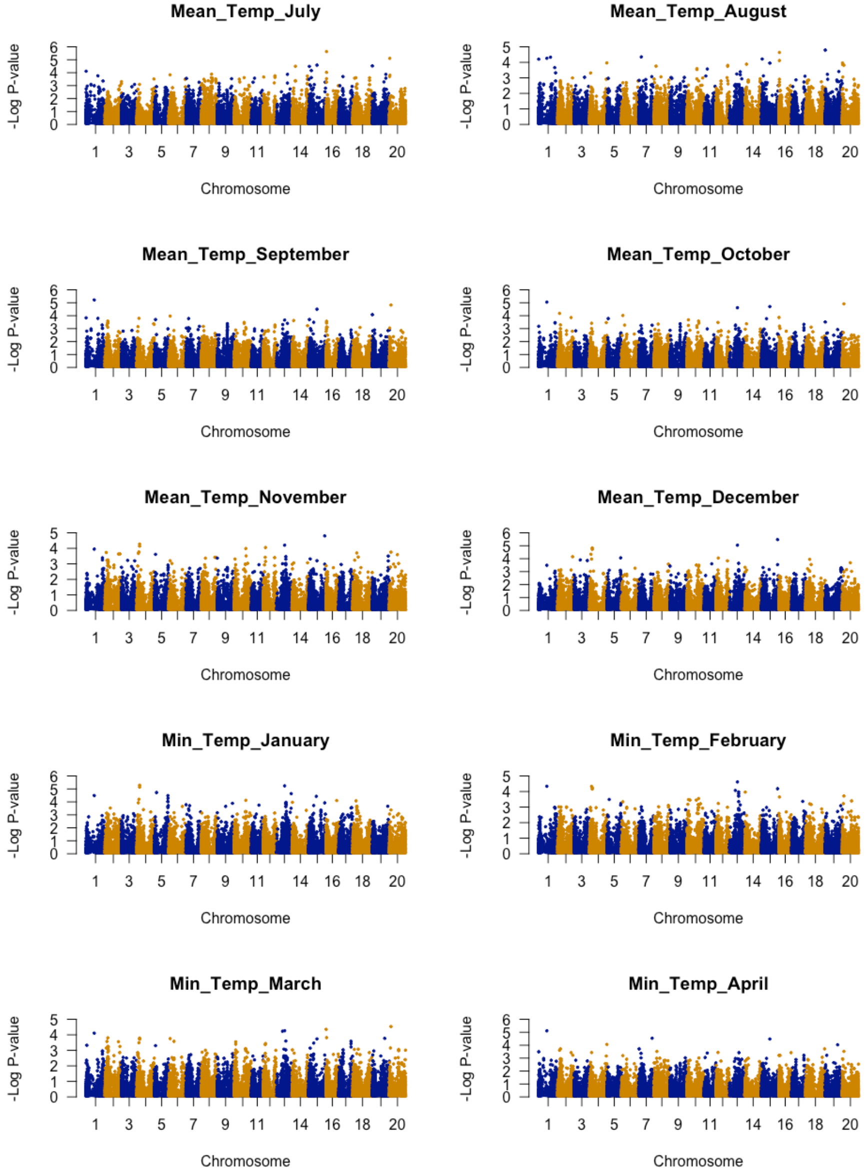


Mean_Temp_May



Mean_Temp_June




Mean_Temp_July

Mean_Temp_August

Mean_Temp_September

Mean_Temp_October

Mean_Temp_November

Mean_Temp_December

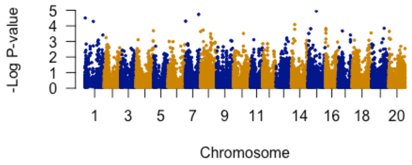
Min_Temp_January

Min_Temp_February

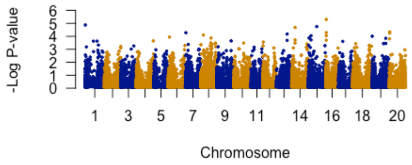
Min_Temp_March

Min_Temp_April

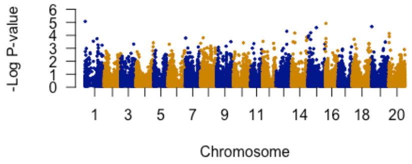
Min_Temp_May



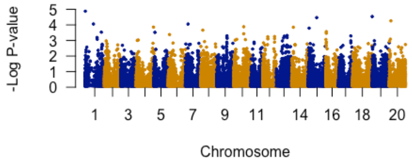
Min_Temp_June



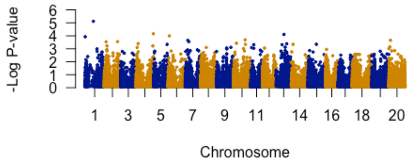
Min_Temp_July



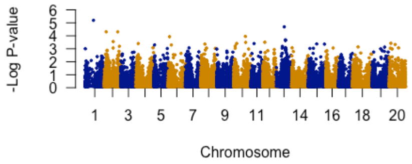
Min_Temp_August



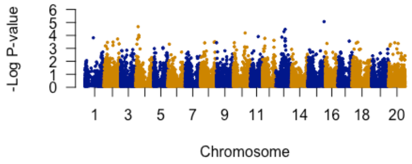
Min_Temp_September



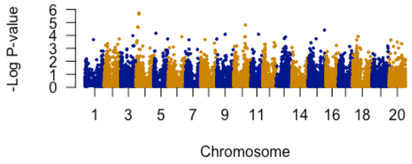
Min_Temp_October



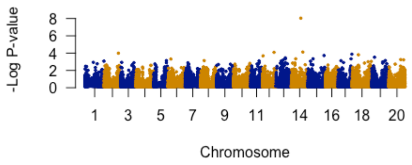
Min_Temp_November



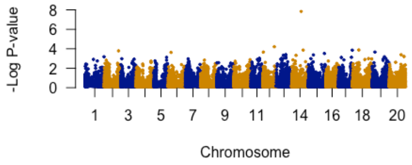
Min_Temp_December

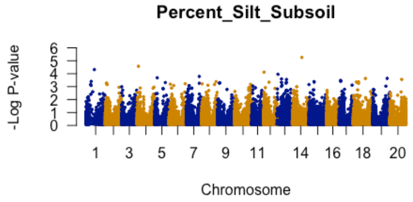


Percent_Sand_Subsoil

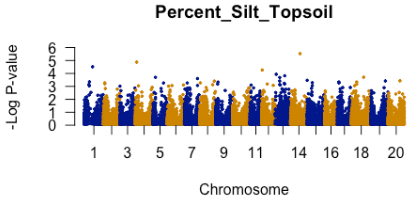


Percent_Sand_Topsoil

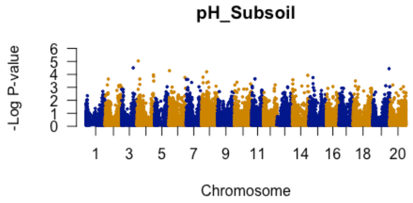




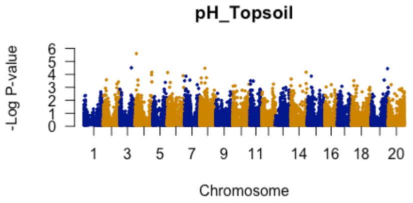
Percent_Silt_Subsoil



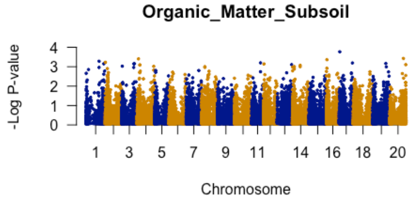
Percent_Silt_Topsoil



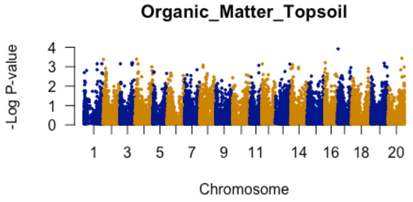
pH_Subsoil



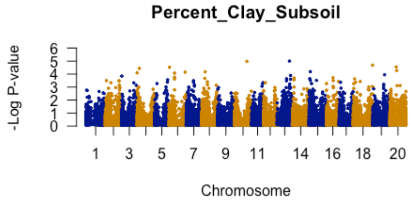
pH_Topsoil



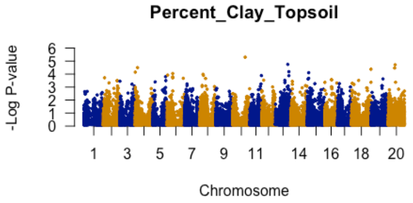
Organic_Matter_Subsoil



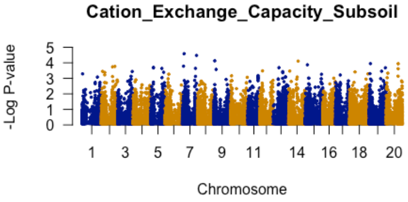
Organic_Matter_Topsoil



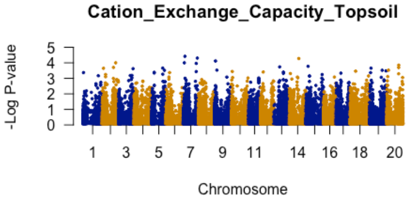
Percent_Clay_Subsoil



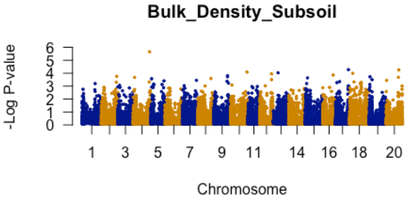
Percent_Clay_Topsoil



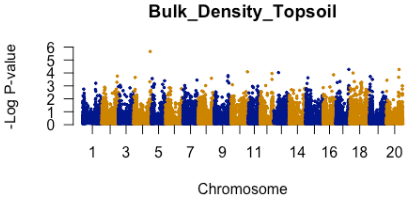
Cation_Exchange_Capacity_Subsoil



Cation_Exchange_Capacity_Topsoil



Bulk_Density_Subsoil



Bulk_Density_Topsoil

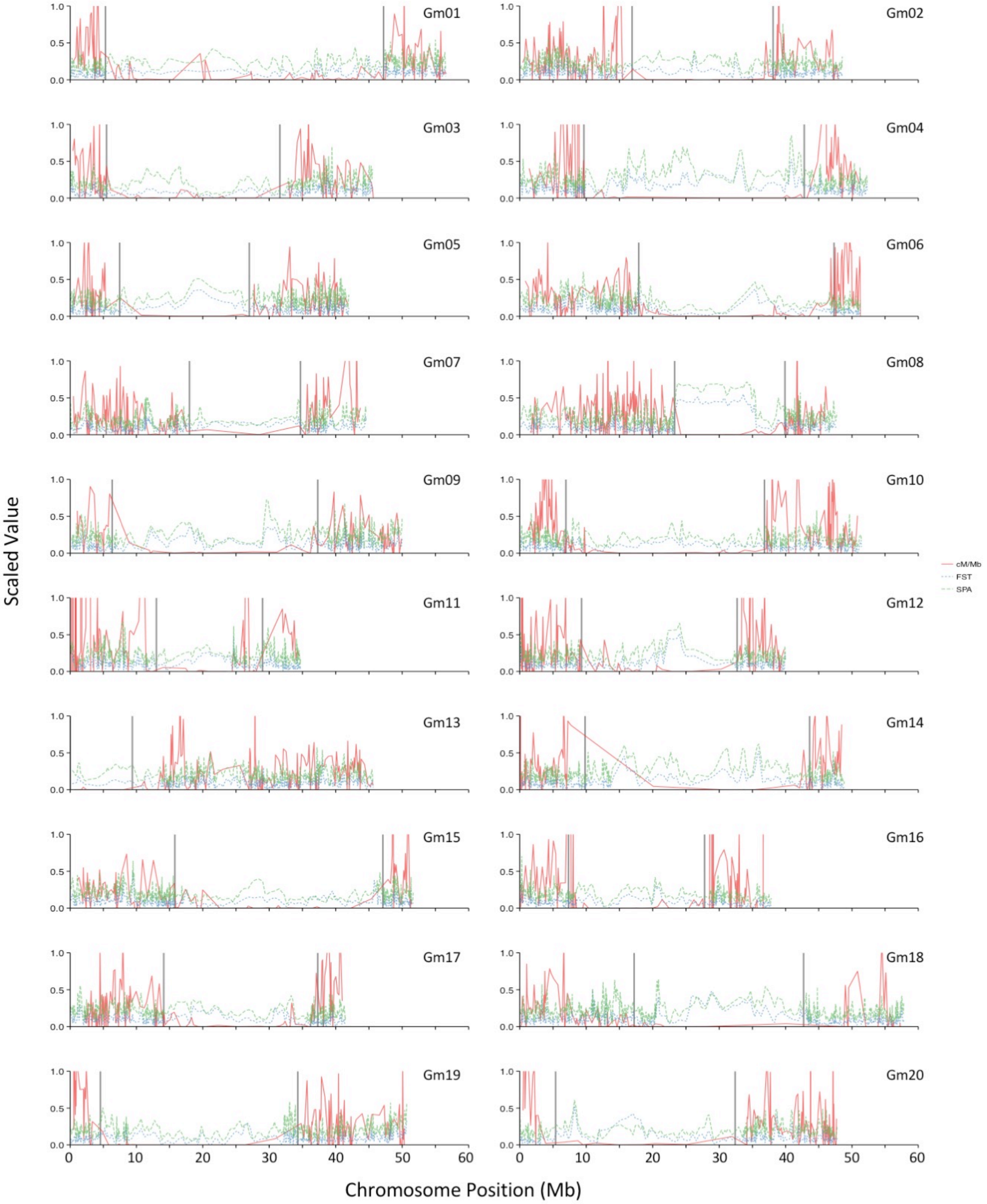


Figure S13. Genome-wide distribution of SPA scores (green dashed line), F_{ST} (blue dotted line), and recombination rate (red solid line). Boarders of pericentromeric regions are denoted with gray vertical lines as annotated on Soybase, corresponding to regions of reduced recombination. SPA values are scaled based on maximum value and plotted based on sliding window average of five markers with a step of three. F_{ST} values are not scaled and plotted based on sliding window average of five markers with a step of three. Recombination rate is scaled on cM/Mb divided by 15 and plotting the midpoint of the physical position.

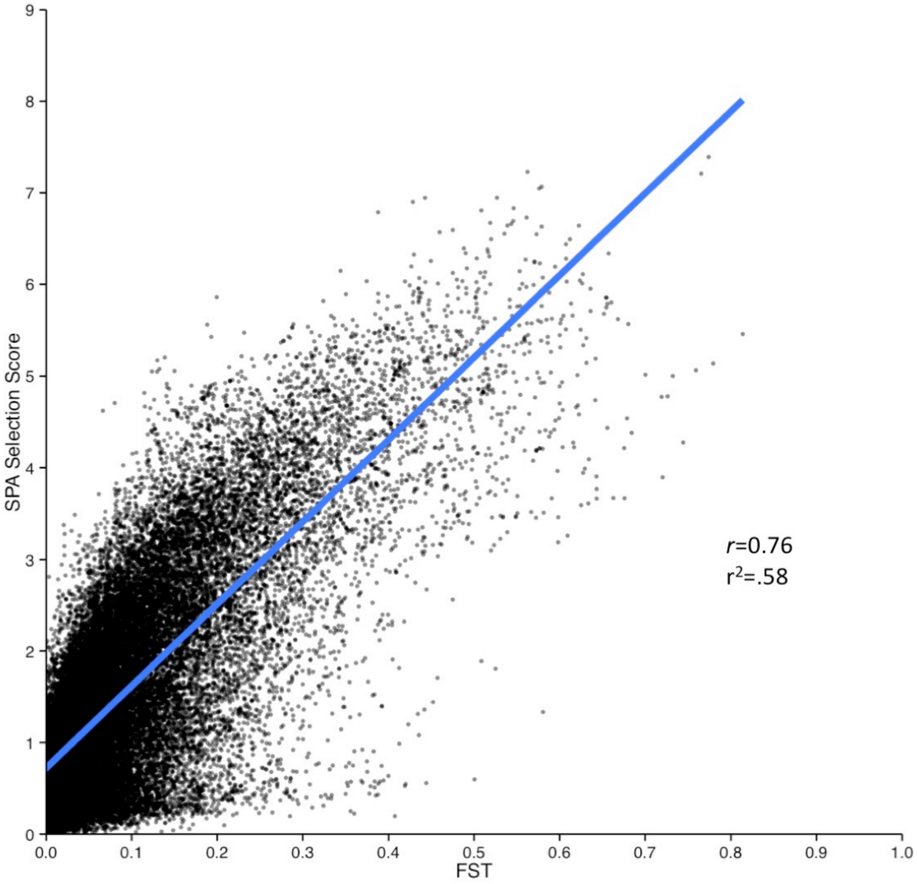
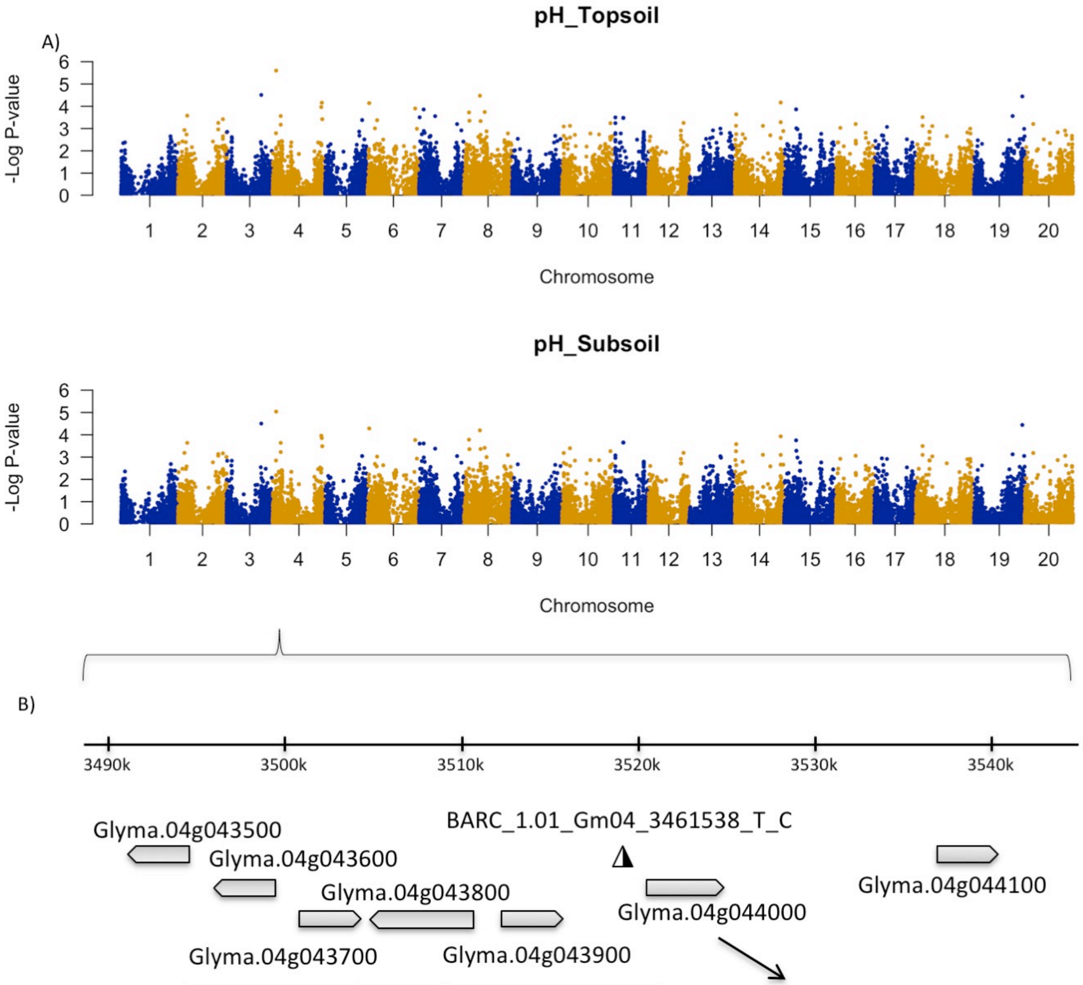


Figure S14. Correlation between SPA and F_{ST} scores.



C)

Study:	533 lines	Song et al., 2013		
Genotypic Class:	G. soja	Landrace	Elite	G. soja
"T" Allele Freq:	0.080	0.063	0.895	0.231

AT1G47240, NRAMP2
Metal ion transmembrane transporter activity

Figure S15. A) Genome-wide associations with topsoil pH and subsoil pH. B) Zoom in on 60 kb region around the significant marker BARC_1.01_Gm04_3461538_T_C. The Arabidopsis homolog for the nearest gene, Glyma.04g044000, is AT1G47240, NRAMP2 Metal ion transmembrane transporter activity (Lanquar et al., 2005). C) The "T" allele is common in *G. soja,* rare in landraces, but frequent in elite lines (Song et al., 2013).

Appendix 3: Publication List

Publications (Peer-reviewed)

Bolon, Y. T., Stec, A. O., Michno, J. M., Roessler, J., Bhaskar, P. B., Ries, L., Dobbels, A. A., Campbell, B. W., Young, N. P., **Anderson, J. E**., ... & Stupar, R. M. (2014). Genome Resilience and Prevalence of Segmental Duplications Following Fast Neutron Irradiation of Soybean. *Genetics*, genetics-114.

Anderson, J. E., Kantar, M. B., Kono, T. Y., Fu, F., Stec, A. O., Song, Q., ... & Stupar, R. M. (2014). A roadmap for functional structural variants in the soybean genome. *G3: Genes| Genomes| Genetics*, g3-114.

Curtin, S. J., **Anderson, J. E.**, Starker, C. G., Baltes, N. J., Mani, D., Voytas, D. F., & Stupar, R. M. (2013). Targeted Mutagenesis for Functional Analysis of Gene Duplication in Legumes. In *Legume Genomics* (pp. 25-42). Humana Press.

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., **Anderson, J. E.**, Hyten, D. L., ... & Stupar, R. M. (2012). Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes. *Plant physiology*, 159(4), 1295-1308.

Publications (Not peer-reviewed)

Anderson, J., Jacobson, A., Swegarden, H., & Tiede, T. (2014). Empowering graduate students in the plant sciences. CSA News.