The role of deleterious substitutions in crop genomes

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Thomas Kono

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

Peter L. Morrell and Robert M. Stupar

October 2016

Acknowledgements

I would like to thank my advisors, Dr. Peter L. Morrell and Dr. Robert M. Stupar, for providing me with guidance and opportunities to advance my career in their labs. Without their support, I would not have had a successful and fulfilling graduate education. Similarly, I would like to thank my committee members, Dr. Ruth G. Shaw, Dr. Kevin P. Smith, and Dr. Peter Tiffin, for their insight and perspective on the work presented here. They challenged me to approach problems from angles I would not have considered on my own, which greatly enriched my experience in this program. I would additionally like to thank present and past members of the Morrell and Stupar Labs for providing a stimulating environment in which to work.

I would also like to extend gratitude to my collaborators on the work presented here. I could not have done any of this work without a wide team of collaborators. In particular, I would like to thank Dr. Amber Eule-Nashoba, Dr. Justin Fay, Dr. Fengli Fu, Paul Hoffman, Chaochih Liu, and Dr. Mohsen Mohammadi for their contributions to my work. Outside of the work presented in this dissertation, I would like to thank Dr. Justin Anderson, Dr. Benjamin Campbell, Dr. Zhou Fang, Dr. Michael Kantar, Jean Michel-Michno, Dr. Liana Nice, and Dr. Ana Poets for providing opportunities for me to expand my skillset and participate in their research work.

Additionally, I would like to acknowledge and thank funding sources and institutional facilities for supporting my research and me. My first three years were supported by a United States Department of Agriculture National Needs Fellowship. My fourth year was supported by a MnDRIVE Global Food Ventures fellowship, and my

fifth year by a University of Minnesota Doctoral Dissertation Fellowship. Travel to conferences and workshops were supported by the Department of Agronomy and Plant Genetics in addition to the aforementioned sources. The analyses presented in this dissertation were carried out with hardware and software support from the University of Minnesota Supercomputing Institute.

Thank you to my Applied Plant Sciences graduate student cohort for providing a social support network. I would also like to thank the students of the Plant Biological Sciences program for including me in their functions.

Finally, I am very grateful to my family and friends for supporting me during this degree program. Without you, I would not have the capacity to pursue a career in the sciences.

Abstract

Historically, it has been postulated that populations carry genetic variants with deleterious effects as segregating variation. Recent advances in DNA resequencing technology allow a new view on this topic, providing a way to identify segregating molecular sequence variants that alter sites that show constrained evolution. Targeted identification and removal of such segregating variants has been proposed as a novel path for improving agronomic performance in plant breeding programs. In this dissertation, I present a brief review of deleterious variation, and why it may be important to plant breeding. Then, I present a tool for classifying single nucleotide polymorphisms into functional classes using public sequence databases. Next, I present a survey of potentially deleterious alleles segregating in two crop species, and a software package that implements a likelihood ratio test for sequence constraint and deleterious prediction. Finally, I present an examination of the contribution of potentially deleterious alleles to yield in a barley breeding experimental population. While potentially deleterious alleles themselves do not explain a larger proportion of phenotypic variation than non-deleterious alleles, genomic prediction of phenotypes may be improved by examining informative subsets of sequence variants, rather than with genomewide markers.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The role of mutation in contributing deleterious alleles to populations and the factors that permit these variants to segregate in populations are classical issues in evolutionary biology. A mutation is said to be deleterious if it has negative impacts on fitness. Fisher, in *Genetical Theory of Natural Selection*, argues that species are generally well-adapted to the environments in which they reside, and random mutation is almost always deleterious (Fisher 1930). Dobzhansky addressed similar topics in *Genetics and the Origin of Species*, in which he recognizes that the segregation of deleterious alleles is a necessary consequence of genetic variation in a population, and may allow a population to respond to a change in environmental conditions (Dobzhansky 1937). Discussions surrounding the segregation of deleterious alleles eventually led to the development of the concept of genetic load (Muller 1950). Briefly, genetic load is a population concept, and is defined as the difference between population mean fitness and the fitness of a hypothetical optimal genotype (Crow 1958). It arises due to new mutations, heterozygote advantage, or "mismatch" between genotypes and environments. It has been shown mathematically (Kimura et al. 1963) and in computer simulation (Felsenstein 1974) that deleterious alleles accumulate in finite populations and contribute to reductions in absolute fitness.

Discussions of genetic load mostly surround the interpretability of a comparison to a hypothetical optimal genotype (Wallace 1970). The numerical value of genetic load in a population depends on the reference genotype used to estimate the optimal fitness value. In the case of recurrent deleterious mutation in a monomorphic population from the

optimum genotype to an unfavorable genotype, the genetic load depends on the mutation rate to the deleterious allele, which can be numerically small (Crow 1958). However, if a beneficial mutation arises in that same population, then the genetic load becomes orders of magnitude larger, as every un-mutated allele constitutes genetic load (Sangvhi 1963). The absolute fitness values for each genotype remains constant, but the genetic load depends on the point of reference, meaning that enumeration of genetic load may not be relevant for the evolution of populations. Additionally, the genetic load equations assume that a population is in mutation-selection equilibrium, and that alleles do not have conditional fitness effects. While non-equilibrium conditions and fluctuating selective coefficients were outside the scope of the original equations (Crow 1963), they are factors in the evolution of real populations, and represent a limitation of applying genetic load concepts to the study of real populations.

Recent advances in DNA sequencing technology allow for identification of genome-wide molecular variants segregating in populations. Each segregating variant must necessarily have an associated selective effect, ranging from lethal to neutral to advantageous (Eyre-Walker and Keightley 2007). The expression of the fitness effect of any given variant will depend on the environmental condition in which it is evaluated. While the selective effect of an individual molecular variant is nearly impossible to measure (Thatcher et al. 1998), segregation of all variants in bulk may explain a substantial portion of phenotypic variation (Yang et al. 2010). In fact, limitations to quantitative trait locus mapping and association analyses (Rockman 2012) have led to a paradigm shift in animal and plant breeding. The genomic prediction and selection

strategies used in breeding programs draw from the ability to predict phenotypic values from genome-wide molecular variants, rather than attempting to associate a single variant with a phenotypic impact (Meuwissen et al. 2001).

There have also been improvements in approaches that predict the functional impact of a variant based on sequence constraint. These approaches are mostly frequently applied to amino acid sequence, as noncoding sequence is typically more difficult to align owing to lower phylogenetic sequence constraint (Graur and Li 2000). Most of the approaches that make use of sequence conservation are simple heuristics - in an alignment of related sequences, a variant that disrupts a conserved amino acid residue is predicted to disrupt protein function. Variants that are predicted to disrupt protein function are then assumed to be deleterious. Most tools of this type only predict the impact of variants segregating in human populations. However, two of the most commonly used programs, Sorting Intolerant from Tolerated (SIFT) (Ng 2003) and Polymorphism Phenotyping 2 (PPH2) (Adzhubei et al. 2013), are able to predict impacts in other systems by using public sequence databases as a source for phylogenetic constraint estimation.

When identifying molecular variants from DNA resequencing data, several forms of bias must be addressed. Reference bias, a special form of ascertainment bias, has effects on both variant discovery and deleterious predictions. Resequenced samples are generally compared to a reference assembly, and samples that are more closely related to the reference strain are inferred to have fewer sequence variants. This results in those samples having fewer putatively deleterious variants. Further, when predicting whether or

not a variant is deleterious, inclusion of the reference assembly for the query species biases the predictions. At sites where the reference carries the derived allele, the variant is systematically predicted to be not deleterious (Simons et al. 2014).

Another source of bias in predicting deleterious variants arises from using a single set of annotated gene models for SNP classification. Predicting a variant as deleterious depends on its assigned amino acid sequence impact. This depends on the gene models that are annotated on the reference assembly. Populations are known to contain variation for transcription start position, stop position, and splicing forms of messenger RNA (Gan et al. 2011). Additionally, the technical issues associated with the quantitative and conditional nature of transcript abundance affect which gene models that are annotated in the reference assembly. Low abundance or conditionally expressed transcripts may be undiscovered, and variants in them cannot be predicted.

Overcoming these forms of bias can be challenging. To address bias due to gene model annotations, a catalogue of the transcript variation in a population is necessary (Gan et al. 2011; Hirsch et al. 2014). To address the issue of reference bias, data from a suitable outgroup must be available. An ideal outgroup is one that has been isolated from the population of interest long enough such that the two groups show reciprocal monophyly at all genetic loci (Rosenberg 2003). With such an outgroup, polarization of identified variants into ancestral and derived states is unambiguous, and the number of deleterious alleles an individual carries does not depend on the reference strain.

With the combination of recently technological and methodological advances, it is now possible to predict which molecular sequence variants may be deleterious, and test

whether or not they contribute more to phenotypic variation than putatively neutral variants. In this dissertation, I first report a program to gather metadata about single nucleotide polymorphism (SNP) markers using public sequence databases. The program, SNPMeta, allows SNP markers to be binned according to how they alter protein coding sequence. I then report a survey of putatively deleterious SNPs segregating in the protein coding regions of barley (*Hordeum vulgare* ssp. *vulgare*) and soybean (*Glycine max*). I give circumstantial evidence that SNPs identified as deleterious through sequence conservation approaches may be truly deleterious in these species. Finally, I present an estimate of the contributions of putatively deleterious SNPs to grain yield and disease severity using data from a barley breeding experiment. Based on this work, much has been learned about the challenges and limitations in identifying potentially deleterious variants in DNA resequencing data. While the final results do not directly support suggested directions for purging of deleterious variants in plant breeding programs (Morrell et al. 2011; Mezmouk and Ross-Ibarra 2014), important principles for analyzing DNA resequencing data and potential directions for genomic prediction come from this work.

When examining the contributions of putatively deleterious variants to grain yield in a barley breeding program, I find that putatively deleterious SNPs explain slightly more phenotypic variation on average than putatively neutral SNPs. This is result is similar to findings published by Edwards et al. (2016), which show that prediction accuracy for a quantitative trait is increased by enriching the set of markers used for prediction for variants in biological pathways known to be associated with that

quantitative trait. Future genomic prediction and selection efforts may be improved by genotyping informative subsets of markers, rather than sampling random genome-wide markers.

Chapter 2: SNPMeta: SNP annotation and SNP metadata collection without a reference genome[1]

The increase in availability of resequencing data is greatly accelerating single nucleotide polymorphism (SNP) discovery and has facilitated the development of SNP genotyping assays. This, in turn, is increasing interest in annotation of individual SNPs. Currently, annotations are only available through curation, or comparison to a reference genome. Many species lack a reference genome, but are still important genetic models or are significant species in agricultural production or natural ecosystems. For these species, it is possible to annotate SNPs through comparison to cDNA, or data from well annotated genes in public repositories. We present SNPMeta, a tool which gathers information about SNPs by comparison to sequences present in GenBank databases. SNPMeta can annotate SNPs from contextual sequence in SNP assay designs and SNPs discovered through genotyping by sequencing (GBS) approaches. SNPMeta can therefore be used to annotate SNPs in non-model species, or species that lack a reference genome. Annotations generated by SNPMeta are highly concordant with annotations obtained from a reference genome.[1]

---

[1] This work was published in *Molecular Ecology Resources* in November 2013, with full citation information given below. This work was a collaborative effort, with several authors contributing to the final manuscript: the annotation workflow was designed by TJYK, KS, and PLM. The code for implementing the workflow was written by TJYK. Testing for accuracy was performed by KS. Additional testing datasets were provided by JAP. The manuscript was written by TJYK, KS, and PLM. All authors read and approved the manuscript.

Kono TJY, Seth K, Poland JA, Morrell PL. 2014. SNPMeta: SNP annotation and SNP metadata collection without a reference genome. *Mol Ecol Resour*. 14:419-425.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most abundant class of genetic variation within individual genomes, and are widely used as genetic markers. Along with nucleotide state, each SNP can potentially be associated with additional information such as genomic location and contribution to changes in protein sequence. This additional information about SNPs (metadata) becomes more relevant as the number of markers increases: surveying an increased number of SNPs increases the probability of querying a causative mutation underlying a phenotype. Extensive resequencing in humans has led to the discovery of large numbers of genetic polymorphisms and accumulation of extensive catalogs of mutations putatively associated with human disease and phenotypic variation (Amberger et al. 2009). Large volumes of resequencing data in other species such as Drosophila (King et al. 2012) and Arabidopsis (Weigel and Mott 2009) have begun to be used to discover variation underlying phenotypes.

Online databases such as the National Center for Biotechnology Information (NCBI) resource dbSNP (Sherry et al. 2001) or organism-specific online sites such as the maize genetics resource Panzea (http://www.panzea.org/) store and provide both SNP contextual sequence and SNP metadata. However, the depth of the information available is limited to that provided by the submitter. This is not an issue for organisms with finished and well-annotated genome assemblies; tools such as ANNOVAR (Wang et al. 2010) and SnpEff (Cingolani et al. 2012) have been designed to use a reference genome

to gather SNP metadata. This option is not available for organisms where a reference genome is not available and where a finished reference genome may remain impractical due to expense or limitations of current sequencing technology, such as read length and assembly tools. Furthermore, many genome assemblies are incomplete, due to genome size or complexity (Morrell et al. 2012). A number of species, such as sunflower (Kane et al. 2011), sugarcane (Dillon et al. 2007) and wheat, are economically important despite the lack of a reference genome. In systems such as these, SNP metadata may still be available through comparison to genomes of closely related species or to annotated genes and cDNA resources. We present a tool, SNPMeta, to annotate SNPs against publicly available sequence data in the absence of an annotated reference.

MATERIALS AND METHODS

Description and Features

SNPMeta handles all the steps necessary to produce SNP annotations. SNPMeta makes use of SNP contextual sequence from either SNP assays (Gunderson et al. 2005) or genotyping by sequencing (GBS) (Elshire et al. 2011) and returns a finished report. SNPMeta reports are suitable for submission to dbSNP or organism-specific databases, such as the T3 database (Blake et al. 2012), which stores phenotype and genetic marker data for barley and wheat breeding. SNPMeta uses BLAST (Zhang et al. 2004) to identify the GenBank record that best matches a SNP contextual sequence, and then uses alignment tools in the EMBOSS (Rice et al. 2000) software package to position the SNP in the GenBank record. It uses the BioPython library (Cock et al. 2009) to parse the

record and alignment, and gathering metadata for each SNP. The information returned includes the gene name in which the SNP is found (if the SNP is genic), whether the SNP is coding or noncoding, and whether the SNP is synonymous or nonsynonymous. If the SNP is noncoding, SNPMeta reports the distance to the closest annotated coding sequence. In many cases, GenBank records originate from the target species or a closely related species. A summary of the process is shown in Figure 1; more details are available in Appendix S1 and Figure S1. The user manual, including a brief tutorial, is available in Appendix S2.

BLAST Settings

By default SNPMeta uses the BLAST program 'blastn,' accepts a maximum of five matches, and accepts hits below an e-value of 5×10^{-10}. Default settings were determined by manual annotation of several hundred Barley Oligonucleotide Pool Assay (BOPA) SNPs from Close et al. (2009). 'Blastn' performs a nucleotide-to-nucleotide search of the nonredundant (nr) sequence database, and the e-value threshold typically limits results to sequences originating from organisms in the same family.

Along with 'blastn,' SNPMeta can perform searches with 'tblastx,' which compares a six-frame translation of the SNP contextual sequence against a six-frame translation of the sequence database. Using this BLAST program allows matches from slightly more divergent taxa, especially within coding regions, because protein sequences tend to be more conserved than the nucleotide sequences (Graur and Li 2000). Users are able to modify the number of BLAST hits returned, and the maximum e-value of the hits. Increasing the number of hits returned by BLAST increases the chances of producing an

annotation, but only if related species have ample genetic resources. Increasing the e-value threshold also increases the number of annotations produced, at the cost of reduced annotation accuracy due to annotating from more distantly related species.

Technical Data and Input Formats

SNPMeta was developed using Python version 2.7.3, BioPython version 1.59, and EMBOSS version 6.5.7 and tested using Apple OS X 10.8.4. Annotations used in this paper were generated on Linux (CentOS 6.3, 64-bit SMP) with Python version 2.7.2, BioPython version 1.58 and EMBOSS version 6.2.0. SNPMeta has also been tested on older versions of OS X and Microsoft Windows. A full list of test platforms can be found in Appendix S2.

SNPMeta accepts DNA sequences in the FASTA format, with SNPs indicated by IUPAC ambiguity codes. The input files must be plain text, with no other formatting. SNPMeta can output a format suitable for dbSNP submission or a tab-delimited text file. Details of the input file format and the output file format can be found in Appendix S2.

Testing

GBS Data

GBS approaches use high throughput sequencing to both identify and score SNPs in populations. Most rely on a reduction of the genome, involving the use of restriction enzymes to create a sequencing library (Baird et al. 2008; Davey et al. 2011; Elshire et al. 2011; Poland et al. 2012). Recent implementations of GBS make use of enzymes that cut more frequently, which results in more even spacing of markers across the genome (Andolfatto et al. 2011). The end product of a GBS pipeline is a large number of SNPs

distributed across the genome, each embedded in contextual sequence derived from short-read sequencing.

To test SNPMeta with GBS data, we annotated SNPs from three different GBS datasets. First, we used restriction-site associated sequencing (Baird et al. 2008) of the Drosophila Synthetic Population from King et al. (2012) to identify and annotate SNPs against *Drosophila melanogaster* and *D. simulans*. There are multiple Drosophila species with finished reference genomes, so it is possible to both verify SNPMeta's inferred annotations and compare annotation relative to a divergent species. To test SNPMeta annotation of GBS datasets from species with limited genomic resources, we used 5,000 GBS SNPs from both hexaploid oat (*Avena sativa*) and Tausch's goatgrass (*Aegilops tauschii*), and 13,649 GBS SNPs from barley (*Hordeum vulgare* ssp. *vulgare*). All SNPs were identified using the GBS approach of Poland et al. (2012). The oat and goatgrass SNPs were a random sample of 5,000 SNPs from a larger dataset, and the barley SNPs were from GBS of the cultivar Morex, for which the published draft reference genome is available (Mayer et al. 2012). SNPs from these species can potentially be annotated based on public data from *Brachypodium distachyon* and barley. Finally, we tested annotation of 5,000 SNPs each from cotton (*Gossypium barbadense*), zebrafish (*Danio rerio*), dog (*Canis lupus*), chicken (*Gallus gallus*), and horse (*Equus caballus*) from dbSNP. For testing purposes, we focused on biallelic SNPs with a high reported genotyping success rate.

To identify SNPs in *D. melanogaster*, we used Bowtie 2 (Langmead and Salzberg 2012) to map reads from three recombinant inbred lines (12001, 12105, and 21001 in

King et al., 2012) to the *D. melanogaster* genome, release 5.41 from FlyBase (McQuilton et al. 2012). We chose release 5.41 because the GenBank sequences are based on this release. We then used SAMtools mpileup (Li et al. 2009) to call SNPs. We identified a total of 9,766 SNPs, somewhat evenly distributed among the three lines. Using the variant call format (VCF) file from SAMtools and the reference sequence, we generated contextual sequence of 60bp for each SNP, formatted to the guidelines described in the SNPMeta user manual (Appendix S2). These FASTA files were then annotated with SNPMeta, using the default settings. To increase speed and reliability of BLAST searches, we used a local copy of the 'nt' database, current as of October 16, 2012.

To test annotation against a divergent genome, we annotated the same SNPs, but conditioned on the records originating from the related species *D. simulans*. *Drosophila melanogaster* and *D. simulans* have estimated divergence of ~11 million generations (Cutter 2008), and synonymous site divergence between *D. melanogaster* and *D. simulans* is ~11% (Begun et al. 2007). We limited the list of BLAST hits using an Entrez query, but it did not completely filter the results to *D. simulans*. We ensured annotation against *D. simulans* by requesting 25 hits per search (instead of the default five), and conditioning on GenBank records originating from *D. simulans*.

For GBS SNPs, a custom Python script was used to convert the GBS tags into a format suitable for SNPMeta input; this code is included with the SNPMeta distribution. The goatgrass SNPs were annotated with default settings in SNPMeta. Since oat does not have many close relatives with well-developed genomic resources, the oat SNPs were annotated with 'tblastx' and an E-value threshold of 5×10^{-5}, instead of the default $5\times10^{-}$

[10], to attempt to annotate as many SNPs as possible. The SNPs downloaded from dbSNP were also annotated with default settings.

Illumina Contextual Sequence Data

To test SNPMeta against data from SNP assays, we annotated SNPs from Illumina SNP assays in both barley and wheat. The main difference between these SNP assays and SNPs discovered through GBS approaches is that these SNP assays were primarily designed from expressed sequence tags (ESTs) and PCR amplicons (Close et al. 2009; Cavanagh et al. 2013). The wheat SNPs were discovered by building a set of reference transcripts from nine wheat cultivars, and assembling short reads against those (Cavanagh et al. 2013). In total, our wheat dataset comprised 8,632 SNPs. The barley SNPs were composed of two production-scale GoldenGate assays, called Barley Oligonucleotide Pool Assay 1 (BOPA1) and BOPA2 (Close et al. 2009). Additionally, we annotated 1,523 SNPs from the pilot OPA panels, and 5,010 barley SNPs from The James Hutton Institute (http://bioinf.scri.ac.uk/barley_snpdb/). In total, our barley dataset contained 9,606 SNPs.

RESULTS

GBS Data

When annotating from the best GenBank match, SNPMeta was able to generate annotation information for 7,636 (78.2%) of the SNPs identified in the three D. melanogaster lines. Of all annotations produced, 7,059 were derived from D. melanogaster. We compared the annotations from SNPMeta by comparing directly to the reference sequence and its accompanying Generic Feature Format v. 3 (GFF) file. Among

SNPs that could be annotated, SNPMeta's annotations matched those from the reference genome for 7,203 (94.3%) SNPs across the three lines. The concordance between SNPMeta's annotations and the GFF annotations for non-coding, synonymous and nonsynonymous variants is shown in Table 1.

Performance was lower when annotating against a divergent species; when conditioning on annotations originating from *D. simulans*, SNPMeta returned annotation for 2,699 (27.6%) SNPs. Additionally, a greater proportion of SNPs was inferred to be nonsynonymous, indicating a potential for a higher error rate when annotating against a divergent species. Comparisons between annotations from *D. melanogaster* and *D. simulans* are summarized in Table 2.

SNPMeta generated annotations for 1,443 (10.6%) of the barley GBS SNPs. Most of the annotations were derived from two barley full-length cDNA libraries (Sato et al. 2009; Matsumoto et al. 2011). SNPMeta generated annotations for 153 (3.06%) of the 5,000 goatgrass GBS SNPs. Twenty-two of the annotated SNPs from goatgrass matched named genes in barley or bread wheat. SNPMeta provided annotation for 66 (1.32%) of the oat SNPs. These annotations were primarily derived from *Brachypodium distachyon*, and include 13 predicted genes in the genome assembly. To remove low confidence SNP annotations, we filtered on the score of the alignment of the SNP contextual sequence to the GenBank record. Since alignment score is dependent on sequence length, we divided the raw alignment score by the contextual sequence length to obtain a "per-base pair" alignment score. Any annotation based on an alignment that scored lower than 4.0 per base pair was removed. The cotton and horse SNPs, in particular, suffered from this

filtering; most of the alignments between the SNP contextual sequence and the GenBank record were of poor quality, but the SNPs that annotated are more likely to be correct. The results from annotation of these species, including the SNPs downloaded from dbSNP, are summarized in Table 3.

When annotating with BLAST reports generated prior to annotation with SNPMeta, the program took 2.5 hours (average 2.8 seconds per SNP) to annotate all 3,232 SNPs called in *Drosophila melanogaster* line '12001.' When running BLAST from within SNPMeta, it takes 115 hours (average 2 minutes and 8 seconds per SNP) to annotate the same SNPs.

Illumina Contextual Sequence Data

SNPMeta annotated 5,377 (62.3%) of the wheat SNPs. Most of the annotations originated from the barley full-length cDNA libraries mentioned above (Sato et al. 2009; Matsumoto et al. 2011). Three-hundred and thirty-two of the annotated SNPs occurred in named genes, primarily from barley and wheat. Annotating against the best GenBank match, SNPMeta was able to annotate 7,774 (80.9%) of the barley SNPs. Of the SNPs that successfully annotated, 6,392 (84.5%) were based on the same full-length cDNA libraries mentioned above (Sato et al. 2009; Matsumoto et al. 2011). Of the annotated SNPs, 5,870 (77.6%) were inferred to be synonymous, and 1,693 (22.4%) were nonsynonymous. Annotations from barley and wheat are also summarized in Table 3.

To assess the accuracy of SNPMeta annotations, we compared SNPMeta annotations to both the high and low confidence gene sets for the draft barley reference genome (Mayer et al. 2012). We used Bowtie 2 (Langmead and Salzberg 2012) to map

the SNP contextual sequences to the draft genome. Of our 9,606 barley SNPs, 5,984 could be mapped to the draft genome and also had SNPMeta annotation information. A custom Python script was then used to see whether each SNP was genic or non-genic in both the high confidence and low confidence gene sets. Concordance between SNPMeta's results and the GFF results exceeded 94% (Table 4). For SNPs inferred to be genic in the high-confidence gene set, concordance between SNPMeta inference for synonymous or nonsynonymous and the GFF annotations was 95.1% (2,902 out of 3,051 SNPs).

DISCUSSION

SNPMeta accurately annotates SNPs against records in GenBank when the SNPs and the records originate from the same organism. For the barley Illumina SNP data, the majority of the annotations were derived from the two full-length cDNA libraries (Sato et al. 2009; Matsumoto et al. 2011) in GenBank, demonstrating the utility of these sequence repositories in gathering annotations. Annotations include 23 fields, including the GenBank number for the best annotated BLAST hit, gene name, if the SNP is in a coding region, and the amino acid states altered by the SNP. These data have many downstream applications and provide a ready means of linking to other genetic resources. For example, in barley, the annotations can be readily linked to the estimated linear order of genes, as part of the barley genome zipper project (Mayer et al. 2011).

Another potential application of SNPMeta is to annotate SNPs relative to data from a closely related species. This is demonstrated both in annotating SNPs from *D.*

melanogaster against *D. simulans*, and annotating goatgrass, oat, and wheat SNPs, where annotations derive primarily from barley cDNA. Accuracy in these scenarios was lower than annotation against sequences from the species of origin. This is not surprising, given that *D. melanogaster* and *D. simulans* are about 11% divergent at synonymous sites (Begun et al. 2007), and barley and wheat are about 12% divergent at synonymous sites (Chalupska et al. 2008). However, SNPMeta still produces useful information. While impact on coding function is less likely to be accurate, gene name and distance to the nearest gene are still available. It should be pointed out that our comparisons here are useful for testing, but likely reflect a greater degree of divergence between species than will typically be used for annotation. Particularly for other crop plants and wild relatives, annotations can be derived from more closely related species. For example, synonymous site divergence between annual sunflower (*Helianthus annuus*) and the well-studied wild relative (*Helianthus petiolaris*) is estimated to be ~ 5% (Strasburg et al. 2011).

While SNPMeta can annotate many SNPs, some proportion of SNPs cannot be annotated, and some are annotated incorrectly. The exact number of SNPs that fall into these two categories depends on the organism, and method for SNP discovery. Organisms with well-developed genetic and genomic resources are more likely to annotate correctly, as in the D. melanogaster example. SNPs discovered from coding sequence, for example EST-derived SNPs, such as the barley OPA and wheat SNPs (Cavanagh et al. 2013) also have a higher probability of successful annotation. From our barley GBS dataset, it is expected that only around 10% of the GBS SNPs can be annotated with the cDNA libraries (Sato et al. 2009; Matsumoto et al. 2011). Because we are annotating barley

GBS SNPs against barley genomic resources, we expect that only around 9% of the GBS SNPs will annotate against available cDNA resources. Taking this into account, annotation information was available for about one fifth of the genic goatgrass SNPs, and almost 15% of the genic oat SNPs.

Most of the discrepancies that arise between SNPMeta annotations and annotations produced from the reference genome are due to mismatches between the GenBank sequences and the reference sequence. For instance, almost 4,000 of the annotations from *D. melanogaster* are against sequences other than the reference genome. For the BOPA SNP annotations, most of the discrepancies arise due to annotation differences between the reference sequence and the GenBank sequences. Differences in annotation include 953 cases where genes were not included in the barley high confidence gene set (Mayer et al. 2012), but were identified in the full-length cDNA libraries (Sato et al. 2009; Matsumoto et al. 2011). Of these, 481 were identified as genes in the low confidence gene set of Mayer et al. (2012). While absent in the high confidence gene set, the presence of these genes in both GenBank and the low confidence gene set suggests that they are *bona fide* genes. In such cases, SNPMeta annotations can potentially provide annotations not available from a reference genome.

SNPs can fail to annotate for a number of reasons, including cases where the SNP is not in proximity to genes or coding sequence in the GenBank record. In these cases, SNPMeta will attempt to process a sequence without annotated coding sequence, and return a 'No Annotations' message. Other sources of failure are alignment errors, in which SNPMeta cannot determine the coding state of the SNP. When alignment issues

occur, including a low alignment score, SNPMeta writes a message into the ‘Notes’ field of the output, signaling further investigation may be necessary. Further details of handling problematic cases are discussed in Appendix S1, and the details of SNPMeta output are described in Appendix S2.

When SNPMeta is run without prior BLAST reports it depends on Web service from NCBI; reliability can be improved by splitting the queries into small batches. During times of high server load, requests can either timeout, or be rejected, which causes SNPMeta to quit. SNPMeta is written to handle service timeouts during peak hours, but it cannot be guaranteed that every run will succeed. In annotating the roughly 10,000 SNPs from *D. melanogaster*, we found that dividing into batches of 800 SNPs greatly improved the potential for batches of annotation to run to completion. This was especially true when performing the more demanding annotation against *D. simulans*. For annotating very large sets (in the tens of thousands) of SNPs, it is more practical to perform the BLAST searches separately from annotation, and have SNPMeta annotate from prebuilt BLAST XML reports or from a local copy of the ‘nt’ database. We provide utility scripts to perform both the partitioning of queries and pre-building BLAST reports.

Recent advances in sequencing technology are greatly increasing the rate of SNP discovery. This, in turn, is increasing interest in the potential function of each SNP, as there is a greater chance of identifying functionally important variants, including the causative mutations underlying a phenotype. SNPMeta is a tool to gather information about SNPs using sequences in GenBank. Because it annotates from GenBank, SNPMeta can be used to compile annotation information for SNPs from non-model organisms, or

organisms which currently lack a reference genome.

	Noncoding	Coding		Total
		Synonymous	Nonsynonymous	
SNPMeta Annotations Matching Reference	2,572 (98.0%)	3,624 (91.4%)	1,007 (96.4%)	7,203 (94.3%)
Total According to Reference	2,624	3,967	1,045	7,636

Table 1: Counts of SNPs in three classes whose annotations from SNPMeta agree with annotations based on the *Drosophila melanogaster* reference genome, excluding missing annotations.

Annotation Against Best GenBank Match			Annotation Against *D. simulans*		
Total	Synonymous and Noncoding	Non-synonymous	Total	Synonymous and Noncoding	Non-synonymous
78.2%	88.0%	12.0%	27.6%	76.2%	21.0%

Table 2: Summary of annotation output for *D. melanogaster* SNPs. The numbers in the 'Annotation Against Best GenBank Match' columns summarize a run where SNPMeta was allowed to use the best annotated BLAST hit for each SNP contextual sequence. The numbers in the 'Annotation Against *D. simulans*' columns summarize a run where SNPMeta was restricted annotation originating from *Drosophila simulans*. 'Total' is the average percentage of SNPs that have complete annotation information

Table 3: Summary of annotation results from nine species. The 'Source' column indicates the source of the SNP data that was used for annotation. 'Total' is the total number of SNPs in that species' dataset. 'Survived QC' is the number of SNPs that passed the alignment score filter. 'Annotated' is the number of SNPs that produced annotation output. Percentages for synonymous, noncoding, and nonsynonymous SNPs are calculated relative the total number of SNPs that annotated. The average sequence lengths for each dataset are also shown. Columns may not add up exactly, since alignment issues can cause missing information. SNPs from Illumina assays, which were biased toward genic sequences, annotated much more successfully than other sources of data.

Organism	Source	Total	Survived QC	Annotated	Syn. and NC	Non-synonymous	Avg. Seq. Len.
Barley (*Hordeum vulgare*)	GBS	13,649	13,415	1,209 (9.01%)	852 (70.5%)	344 (28.5%)	63.24bp
Goatgrass (*Aegilops tauschii*)	GBS	5,000	4,958	110 (2.2%)	69 (62.7%)	37 (33.6%)	62.18bp
Oat (*Avena sativa*)	GBS	5,000	5,000	66 (1.3%)	43 (65.2%)	14 (21.6%)	60.19bp
Barley (*Hordeum vulgare*)	Illumina Assay	9,606	8,773	6,599 (75.2%)	5,199 (78.8%)	1,372 (20.8%)	149.22bp
Bread Wheat (*Triticum aestivum*)	Illumina Assay	8,632	7,234	3,986 (55.1%)	2,440 (61.2%)	1,313 (32.9%)	188.18bp
Chicken (*Gallus gallus*)	dbSNP	5,000	4,869	72 (1.5%)	56 (77.8%)	16 (22.2%)	200.66bp
Cotton (*Gossypium barbadense*)	dbSNP	5,000	4,844	13 (0.3%)	6 (46.2%)	7 (53.8%)	237.07bp
Dog (*Canis lupus*)	dbSNP	5,000	5,000	195 (3.9%)	127 (65.1%)	56 (28.7%)	287.27bp
Horse (*Equus caballus*)	dbSNP	5,000	4,689	40 (0.9%)	24 (60.0%)	15 (37.5%)	401.0bp
Zebrafish (*Danio rerio*)	dbSNP	5,000	4,695	1,269 (25.6%)	708 (55.8%)	558 (44.0%)	526.10bp

	Non-coding	Coding	Total
SNPMeta Annotations Matching Reference	2,256 (100%)	3,386 (90.8%)	5,642 (94.3%)
Total According to Reference	2,256	3,728	5,984

Table 4: Counts of SNPMeta annotations for the barley SNPs that agree with genic or non-genic calls based on the barley draft genome. These numbers exclude SNPs that could not be annotated or confidently mapped to the draft genome.

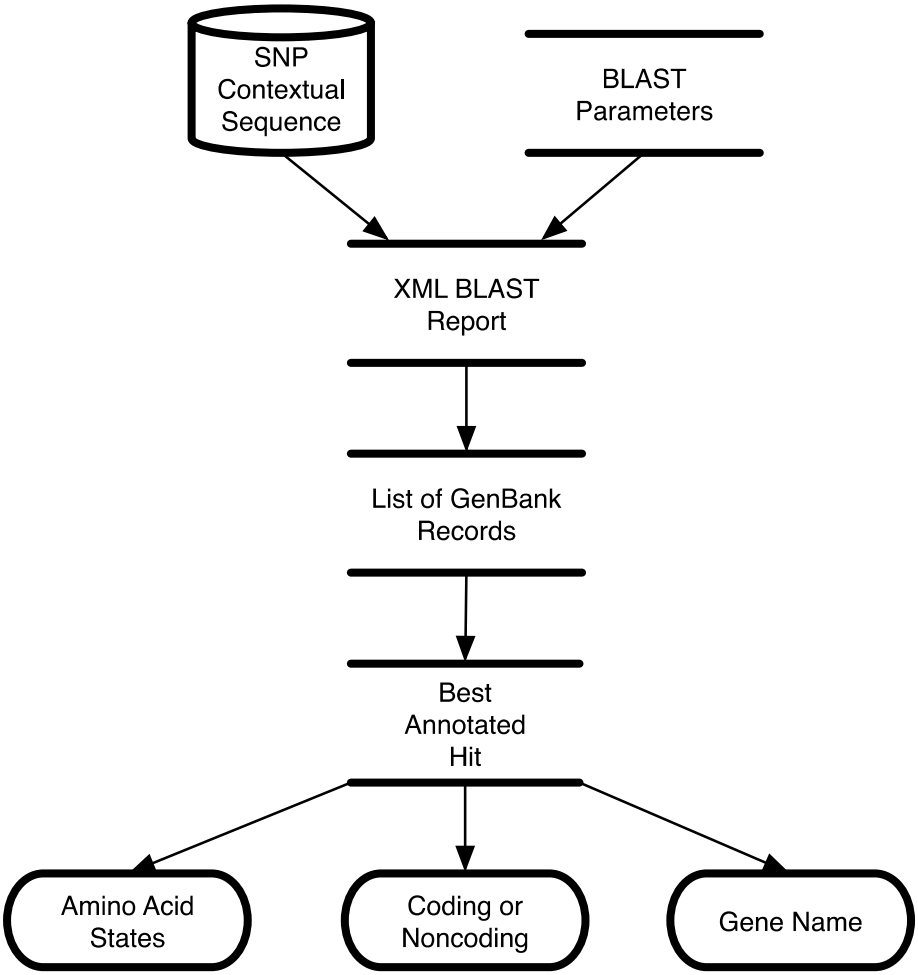

Figure 1: A brief summary of the SNPMeta annotation workflow. SNP contextual sequences are read off disk, and BLAST is used to match the query sequence with annotated records in GenBank. Alignment tools are used to determine coding state. SNPMeta returns the information indicated in rounded boxes.

Chapter 3: The role of deleterious substitutions in crop genomes[2]

Populations continually incur new mutations with fitness effects ranging from lethal to adaptive. While the distribution of fitness effects (DFE) of new mutations is not directly observable, many mutations likely have either no effect on organismal fitness or are deleterious. Historically, it has been hypothesized that a population may carry many mildly deleterious variants as segregating variation, which reduces the mean absolute fitness of the population. Recent advances in sequencing technology and sequence conservation-based metrics for inferring the functional effect of a variant permit examination of the persistence of deleterious variants in populations. The issue of segregating deleterious variation is particularly important for crop improvement, because the demographic history of domestication and breeding allows deleterious variants to persist and reach moderate frequency, potentially reducing crop productivity. In this study, we use exome resequencing of fifteen barley accessions and genome resequencing of eight soybean accessions to investigate the prevalence of deleterious SNPs in the protein-coding regions of the genomes of two crops. We conclude that individual cultivars carry hundreds of deleterious SNPs on average, and that nonsense variants make up a minority of deleterious SNPs. Our approach annotates known phenotype-altering variants as deleterious more frequently than the genome-wide average, suggesting that putatively deleterious variants are likely to affect phenotypic variation. We also report the implementation of a SNP annotation tool (BAD_Mutations) that makes use of a likelihood ratio test based on alignment of all currently publicly available Angiosperm genomes.

[2] This work was published in *Molecular Biology and Evolution* in June 2016, with full citation information given below. This work was a collaborative effort with multiple authors contributing to the final manuscript: the study was designed by TJYK, RMS, PT, KPS, JCF, and PLM. Barley sequence analysis was performed by TJYK and CL. Soybean sequence analysis was performed by TJYK and FF. The deleterious prediction pipeline was developed by TJYK, PJH, and JCF. Information on variants that putatively have phenotypic effects was gathered by TJYK and MM. The manuscript was written by TJYK, JCF, RMS, and PLM. All authors read and approved the final manuscript.

Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. 2016. The role of deleterious substitutions in crop genomes. *Mol Biol Evol*. 33:2307-2317.

INTRODUCTION

Mutation produces a constant influx of genetic variants into populations. Each mutation has a fitness effect that varies from lethal to neutral to advantageous. While the distribution of fitness effects of new mutations is not directly observable (Eyre-Walker and Keightley 2007), most mutations with fitness impacts are deleterious (Keightley and Lynch 2003). It is generally assumed that deleterious mutations alter phylogenetically-conserved sites (Doniger et al. 2008), or cause loss of protein function (Yampolsky et al. 2005). Strongly deleterious mutations (particularly those with dominant effects) are quickly purged from populations by purifying selection. Likewise, strongly advantageous mutations increase in frequency, and ultimately fix due to positive selection (Robertson 1960; Smith and Haigh 1974). Weakly deleterious mutations have the potential to persist in populations and cumulatively contribute significantly to reductions in fitness as segregating deleterious variants (Fay et al. 2001; Eyre-Walker et al. 2006; Doniger et al. 2008).

Considering a single variant in a population, three parameters affect its segregation: the effective population size (N_e), the selective coefficient against homozygous individuals (s), and the dominance coefficient (h). The effects of N_e and s are relatively simple; variants are primarily subject to genetic drift rather than selection if $N_e s < 1$ (Kimura et al. 1963). The effect of h is not as straightforward, as it depends on the genotypic frequencies and the degree of outcrossing in the population. In populations with a high degree of self fertilization or sibling mating, many individuals will be homozygous, which reduces the importance of h in determining the efficacy of selection

against the variant (Glémin 2003). In populations that are closer to panmixia, an individual deleterious variant will occur primarily in the heterozygous state, and *h* will determine how "visible" the variant is to selection, with higher values of *h* increasing the efficacy of selection (Charlesworth and Charlesworth 1999). A completely recessive deleterious variant may remain effectively neutral as long as the frequency of the variant is low enough such that there are not a substantial number of homozygous carriers. Conversely, a completely dominant deleterious variant is expected to be quickly purged from the population (Lande and Schemske 1985). On average, deleterious variants segregating in a population are predicted to be partially recessive (Simmons and Crow 1977), allowing them to remain "hidden" from the action of purifying selection, and reach moderate frequencies. This may be expected, for example, based on data from a gene knockout library in yeast (Shoemaker et al. 1996), which indicate that protein loss-of-function variants have an average dominance coefficient of 0.2 (Agrawal and Whitlock 2012).

Effective recombination rate also has important impacts on the number and distribution of deleterious mutations in the genome. Regions with low effective recombination are prone to the irreversible accumulation of deleterious variants. This phenomenon is known as the "ratchet effect" (Muller 1964). In finite populations with low recombination, the continual input of deleterious mutations and stochastic variation in reproduction causes the loss of individuals with the fewest deleterious variants. Lack of recombination precludes the selective elimination of chromosomal segments carrying deleterious variants, and thus they can irreversibly increase, similar to how a ratchet turns

in only one direction (Muller 1964). Nordborg (2000) demonstrates that under high levels of inbreeding, effective recombination rate can be decreased by almost 20-fold relative to an outbreeding population, showing that mating system can be a major determinant in the segregation of deleterious variation. While inbreeding populations are especially susceptible to ratchet effects on a genome-wide scale, even outbreeding species have genomic regions with limited effective recombination (Arnheim et al. 2003; McMullen et al. 2009). In maize, these low recombination regions are observed to harbor excess heterozygosity in inbred lines, suggesting that they maintain deleterious variants that cannot be made homozygous (Rodgers-Melnick et al. 2015). Both simulation studies (Felsenstein 1974) and empirical investigations in *Drosophila melanogaster* (Campos et al. 2012, 2014) indicate that deleterious variants accumulate in regions of limited recombination.

Efforts to identify deleterious variants and quantify them in individuals have led to a new branch of genomics research. In humans, examination of the contribution of rare deleterious variants to heritable disease has contributed to the emergence of personalized genomics as a field of study (reviewed in Abecasis et al. 2010; Cooper et al. 2010; Marth et al. 2011). Current estimates suggest that an average human may carry ~300 loss-of-function variants (Abecasis et al. 2010; Agrawal and Whitlock 2012) and up to tens of thousands of weakly deleterious variants in coding and functional noncoding regions of the genome (Arbiza et al. 2013). In terms of effects on organismal fitness, the average human carries three lethal equivalents (Gao et al. 2015; Henn et al. 2015). These variants are enriched for mutations that are causative for diseases (Kryukov et al. 2007; Marth et

al. 2011). As such they are expected to have appreciable negative selection coefficients ($N_e s$) and be kept at low frequencies due to the action of purifying selection.

Humans are not unique in harboring substantial numbers of deleterious variants. It is estimated that almost 40% of nonsynonymous variants in *Saccahromyces cerevisiae* have deleterious effects (Doniger et al. 2008) and 20% of nonsynonymous variants in rice (Lu et al. 2006), *Arabidopsis thaliana* (Günther and Schmid 2010), and maize (Mezmouk and Ross-Ibarra 2014) are deleterious. In dogs, Cruz et al. (2008) identified an excess of nonsynonymous single nucleotide polymorphisms (SNPs) segregating in domesticated dogs relative to grey wolves. A similar pattern has been found in horses (Schubert et al. 2014) and sunflowers (Renault and Rieseberg 2015), suggesting that an increased prevalence of deleterious variants may be a "cost of domestication."

Approaches to identify deleterious mutations take one of two forms. Quantitative genetic approaches have been employed that investigate the aggregate impact of potentially deleterious alleles on fitness. Mutation accumulation studies (e.g., Mukai 1964; Schultz et al. 1999; Shaw et al. 2002; Charlesworth et al. 2004) use change in fitness over generations within lineages to estimate mutational effects on fitness. Coupled with DNA sequencing technologies, these studies may shed light on how many DNA sequence changes are potentially deleterious (e.g., Ossowski et al. 2010). On the other hand, purely bioinformatic approaches make use of measures of sequence conservation to identify variants with a significant probability of being deleterious. When combined with genome-scale resequencing, they permit the identification of large numbers of putatively deleterious variants. Commonly applied approaches include SIFT (Sorting Intolerant

From Tolerated) (Ng 2003), PolyPhen2 (Polymorphism Phenotyping) (Adzhubei et al. 2010), and a likelihood ratio test (LRT) (Chun and Fay 2009). These sequence conservation approaches operate in the absence of phenotypic data, but allow assessment of individual sequence variants. As such, some variants identified bioinformatically may be locally adaptive, or conditionally neutral. However, given the observation that deleterious mutations constantly arise and continue to segregate in populations, their targeted identification and elimination from breeding populations presents a novel path for crop improvement (Morrell et al. 2011).

In this study, we investigate the distribution of deleterious variants in thirteen barley (*Hordeum vulgare* ssp. *vulgare*) cultivars, two wild barley (*H. vulgare* ssp. *spontaneum*) accessions, seven soybean (*Glycine max*) cultivars, and one wild soybean (*Glycine soja*) accession using exome and whole genome resequencing. We seek to answer four questions about the presence of deleterious variants: *i*) How many deleterious variants do individual cultivars harbor, and what proportion of these are nonsense (early stop codons) versus nonsynonymous (missense) variants? *ii*) What proportion of nonsynonymous variation is inferred to be deleterious? *iii*) How many known phenotype-altering SNPs are inferred to be deleterious? *iv*) How does the relative frequency of deleterious variants vary with recombination rate? We identify an average of ~1,000 deleterious variants per accession in our barley sample and ~700 deleterious variants per accession in our soybean sample. Approximately 40% of the deleterious variants are private to one individual in both species, suggesting the potential for selection for individuals with a reduced number of deleterious variants. Approximately 3-6% of nonsynonymous variants

are inferred to be deleterious by all three annotation approaches used in our study, and known causative SNPs annotate as deleterious at a much higher proportion than the genomic average. In soybean, where genome-wide recombination rate estimates are available, the proportion of deleterious variants is negatively correlated with recombination rate.

MATERIALS AND METHODS

Plant Material and DNA Sequencing

The exome resequencing data reported here includes thirteen cultivated barleys, and two wild barley accessions. Barley exome capture was based on a 60 Mb liquid-phase Nimblegen capture design (Mascher et al. 2013). For the soybean sample, we resequenced whole genomes of seven soybean cultivars and used previously-generated whole genome sequence of *Glycine soja* (Kim et al. 2010). Each sample was prepared and sequenced with manufacturer protocols (Illumina, San Diego, CA) to at least 25x coverage of the target with 76bp, 100bp or 151bp paired-end reads. A summary of samples and sequencing statistics is given in Table S1.

Read Mapping and SNP Calling

DNA sequence handling followed the "Genome Analysis Tool Kit (GATK) Best Practices" workflow from the Broad Institute (McKenna et al. 2010; DePristo et al. 2011). Our workflow for read mapping and SNP calling is depicted in Figure S1. First, reads were checked for proper length, Phred score distribution, and *k*-mer contamination with FastQC (bioinformatics.babraham.ac.uk/projects/fastqc/). Primer and adapter

sequence contamination was then trimmed from barley reads using Scythe (github.com/vsbuffalo/scythe), using a prior on contamination rate of 0.05. Low-quality bases were then removed with Sickle (github.com/najoshi/sickle), with a minimum average window Phred quality of 25, and window size of 10% of the read length. Soybean reads were trimmed using the fastqc-mcf tool in the ea-utils package (code.google.com/p/ea-utils/). Post-alignment processing and SNP calling were performed with the GATK v. 3.1 (McKenna et al. 2010; DePristo et al. 2011).

Barley reads were aligned to the Morex draft genome sequence (Mayer et al. 2012) using BWA-MEM (Li and Durbin 2009). We tuned the alignment reporting parameter and the gapping parameters to allow ~2% mismatch between the reads and reference sequence, which is roughly equivalent to the highest estimated nucleotide diversity observed at a locus in barley coding sequence (Morrell et al. 2003, 2006, 2014). The resulting SAM file was trimmed of unmapped reads with Samtools (Li et al. 2009), sorted, and trimmed of duplicate reads with Picard tools (picard.sourceforge.net/). We then realigned around indels, using a set of 100 previously known indels from Sanger resequencing of 25 loci (Caldwell et al. 2006; Morrell and Clegg 2007; Morrell et al. 2014). Sequence coverage was estimated with 'bedtools genomecov,' using the regions included in the Nimblegen barley exome capture design (https://sftp.rch.cm/diagnostics/sequencing/nimblegen_annotations/ez_barley_exome/barley_exome.zip). Individual sample alignments were then merged into a multisample alignment for variant calling. A preliminary set of variants was called with the GATK HaplotypeCaller with a heterozygosity (average pairwise diversity) value of 0.008, based

on average coding sequence diversity reported for cultivated barley (Morrell et al. 2014). This preliminary set of variants was filtered to sites with a genotype score of 40 or greater, heterozygous calls in at most two individuals, and read depth of at least five reads. We then used the filtered variants, SNPs identified in the Sanger resequencing data set, and 9,605 SNPs from genotyping assays: 5,010 from the James Hutton Institute (Comadran et al. 2012), and 4,595 from Illumina GoldenGate assays (Close et al. 2009) as input for the GATK VariantRecalibrator to obtain a set of recalibrated variant calls. Final variants were filtered to be supported by a minimum of five reads per sample, have a Phred-scaled genotype quality of at least 40, and have a maximum of two accessions with missing data.

Processing of soybean samples is as described above, but with the following modifications. Soybean reads were aligned to the Williams 82 reference genome sequence (Schmutz et al. 2010). Mismatch and reporting parameters for the cultivated samples were adjusted to allow for ~1% mismatch between reads and reference, which is approximately the highest coding sequence diversity typically observed in soybean (Hyten et al. 2006). The alignments were trimmed and sorted as described above. Preliminary variants were called as in the barley sample, but with a heterozygosity value of 0.001, which is the average nucleotide diversity reported by Hyten et al. (2006). Final variant calls were obtained in the same way as described for the barley sample, using SNPs on the SoySNP50K chip (Song et al. 2013) as known variants.

Transition to transversion ratios were calculated with R scripts. The ratios in the Sanger resequencing dataset were computed using SNPs identified in FASTA alignments

of wild barley gene sequences (Morrell et al. 2006), or a table of SNPs identified in resequencing of soybean gene fragments (supplemental data file 1 in Hyten et al. 2006).

Read mapping scripts, variant calling scripts, and variant filtering scripts for both barley and soybean are available on GitHub at (github.com/MorrellLAB/Deleterious_Mutations).

SNP Classification

Barley SNPs were identified as coding or noncoding using the Generic Feature Format v3 (GFF) file provided with the reference genome (Mayer et al. 2012). A custom Python script was then used to identify coding barley SNPs as synonymous or nonsynonymous. Soybean SNPs were assigned using primary transcripts using the Variant Effect Predictor (VEP) from Ensembl (ensembl.org/info/docs/tools/vep/index.html). Nonsynonymous SNPs were then assessed using SIFT (Ng 2003), PolyPhen2 (Adzhubei et al. 2010) using the 'HumDiv' model, and a likelihood ratio test comparing codon evolution under selective constraint to neutral evolution (Chun and Fay 2009). For the likelihood ratio test, we used the phylogenetic relationships between 37 Angiosperm species based on genic sequence from complete plant genome sequences available through Phytozome (phytozome.jgi.doe.gov/) and Ensembl Plants (plants.ensembl.org/). The LRT is implemented as a Python package we call 'BAD_Mutations' (BLAST Aligned-Deleterious Mutations; github.com/MorrellLAB/BAD_Mutations). Coding sequences from each genome were downloaded and converted into BLAST databases. The coding sequence from the query species was used to identify the best match from each species using TBLASTX. The best

match from each species was then aligned using PASTA (Mirab et al. 2014), a phylogeny-aware alignment tool. The resulting alignment was then used as input to the likelihood ratio test for the affected codon. The LRT was performed on codons with a minimum of 10 species represented in the alignment at the queried codon. Reference sequences were masked from the alignment to reduce the effect of reference bias (Simons et al. 2014). A SNP was identified as deleterious if the p-value for the test was less than 0.05, with a Bonferroni correction applied based on the number of tested codons, and if either the alternate or reference allele was not seen in any of the other species. For barley, our threshold was 8.4E-7 (59,277 codons tested), and for soybean, our threshold was 7.8E-7 (64,087 codons tested). A full list of species names and genome assembly and annotation versions used is available in Table S9.

Relating Recombination Rate to Deleterious Predictions

Recombination rates were taken from a genetic map developed by Lee et al. (2015). Briefly, a recombinant inbred line family was derived from a cross between a wild soybean line and a cultivated soybean line, and genotyped with the SoySNP6K genotyping platform. For our analysis, we calculated cM/Mb values for each interval between markers on the SoySNP6K. Within each interval, we also calculated the proportion of nonsynonymous SNPs that annotated as deleterious by our criteria. Intervals with negative, or cM/Mb values above 20 were excluded, as they indicate regions where the markers likely have incorrect physical position. Pearson correlation (Figure S2A) and logistic regression (Figure S2B) were used to investigate the relationship between recombination rate and deleterious variation.

Inference of Ancestral State

Prediction of deleterious mutations is complicated by reference bias (Chun and Fay 2009; Simons et al. 2014), which manifests in two ways. First, individuals that are closely related to the reference line used for the reference genome will appear to have fewer genetic variants, and thus fewer inferred nonsynonymous and deleterious variants. Second, when the reference strain carries a derived allele at a polymorphic site, that site is generally not predicted to be deleterious (Simons et al. 2014). To address the issue of reference bias, we polarized all coding variants by ancestral and derived state, rather than reference and non-reference state. Ancestral states were inferred for SNPs in gene regions by inferring the majority state in the most closely related clade from the consensus phylogenetic tree for the species included in the LRT. For barley, the ancestral states were inferred from gene alignments of *Aegilops tauschii*, *Brachypodium distachyon*, and T*ritium urartu*. For soybean, ancestral states were inferred using *Medicago truncatula* and *Phaseolus vulgaris*. This approach precludes universal inference of ancestral state for noncoding variants. However, examination of alignments of intergenic sequence in Triticeae species and in *Glycine* species showed that alignments outside of protein coding sequence is not reliable for ancestral state inference (data not shown).

RESULTS

Variant Calling and Identification of Deleterious SNPs

Resequencing and read mapping followed by read de-duplication resulted in an average coverage of ~39X exome coverage for our barley samples and ~38X genome

coverage in soybean. A summary of our resequencing data and read mapping statistics is shown in Table S1. Average heterozygosity was 2.5% in our barley sample, and 0% in our soybean sample, reflecting the inbreeding of the accessions. The observed heterozygosity in our barley sample is mostly due to the inclusion of wild material, which is less inbred than the cultivars. Heterozygous variant calls in soybean were all in reads with low mapping score, possibly due to the highly duplicated nature of the soybean genome (Schmutz et al. 2010). A table of the barley accessions used in this study is shown in Table S2, and the soybean accessions are shown in Table S3.

After realignment and variant recalibration, we identified 652,797 SNPs in thirteen cultivated and two wild barley lines. The majority of these SNPs were noncoding, with 522,863 occurring outside of coding sequence (CDS) annotations. Of the coding SNPs, 70,069 were synonymous, and 59,865 were nonsynonymous. A summary of the variants in various functional classes is shown in Table 1, and a per-sample summary is shown in Table S4. SIFT identified 13,626 SNPs as deleterious, PolyPhen2 identified 13,534 SNPs to be deleterious, and the LRT called 17,865 deleterious. The intersection of all three approaches identifies a much smaller set of deleterious variants, with a total of 4,872 nonsynonymous SNPs identified as deleterious. While individual methods identified ~18% of nonsynonymous variants as deleterious, the intersect of approaches identifies 5.7%. A derived site frequency spectrum (SFS) of synonymous, nonsynonymous, and putatively deleterious SNPs in our barley sample is shown in Figure 1A.

In soybean, we called 586,102 SNPs in gene regions. Of these, 542,558 occured in the flanking regions of a gene model. We identified 73,577 synonymous SNPs, and

99,685 nonsynonymous SNPs (Table S5). SNPs in the various classes sum to greater than the total number of SNPs as a single SNP in multiple transcripts can have multiple functional annotations. For instance, a SNP may be intronic in one transcript, but exonic in an alternative transcript. SIFT identified 7,694 of the nonsynonymous SNPs as deleterious, PolyPhen2 identified 14,933 as deleterious, and the LRT identified 11,223 as deleterious. Per-sample counts of putatively deleterious variants in barley are shown in Table S6, and per-sample counts for soybean are shown in Table S7. Similarly to the barley sample, the proportion of putatively deleterious variants was similar across prediction approaches, with the exception of SIFT, which failed to find alignments for many genes. The overlap of prediction approaches identified 3,041 (2.6%) of nonsynonymous variants to be deleterious (Table 2). Derived allele frequency distributions are shown in Figure 1B. Variants inferred to be deleterious are generally at lower derived allele frequency than other classes of variation, implying that these variants are truly deleterious. For both species, the intersection of approaches appeared to give the most accurate prediction of whether or not a variant is deleterious, as evidenced by enrichment for rare alleles (Figure 2).

Nonsense variants made up a relatively small proportion of putatively deleterious variants. In our barley sample, we identified a total of 711 nonsense variants, 14.5% of our putatively deleterious variants. In soybean, we identified 1,081 nonsense variants, which make up 15.7% of putatively deleterious variants. Nonsense variants have a higher heterozygosity than tolerated, silent, or deleterious missense variants (Figure S3). While the absolute differences in heterozygosity were small due to the inbred nature of our

samples, the pattern suggests that nonsense variants are more strongly deleterious than missense variants.

The transition to transversion ratio in our barley samples was 1.7:1 (Figure S4B), very close to estimates obtained from previous Sanger resequencing in barley genes (Morrell et al. 2006). In soybean, the transition to transversion ratio in our SNPs was 1.4:1, while the estimate from a Sanger resequencing dataset was ~1.2:1 (Hyten et al. 2006). The differences we observe between results from Sanger and Illumina resequencing could be due to the duplicated nature of the soybean genome (Schmutz et al. 2010), leading to paralogous alignment.

Deleterious Mutations and Causative Variants

Bioinformatic approaches to identifying deleterious variants rely on sequence constraint to estimate protein functional impact. An example of a deleterious variant showing a derived base substitution that alters a phylogenetically conserved codon is shown in Figure S5. The variants identified in these approaches should be enriched for variants that cause large phenotypic changes. To explore how frequently known causative SNPs annotate as deleterious, we compiled a list of 23 nonsynonymous variants inferred to contribute to known phenotypic variation in barley and 11 in soybean, and tested the effect of these variants in our prediction pipeline. Of 23 variants that are purported to be causative for large phenotypic changes, 6 (25%) were inferred to be deleterious (Table S8). Of the 11 soybean putatively causative variants, 5 (45%) were inferred to be deleterious. This contrasts with the genome-wide average of ~3-6%, showing that variants our pipeline identifies as deleterious are more likely to impact phenotypes.

Deleterious Mutations and Genetic Map Distance

The effective recombination rate strongly influences purging of deleterious variants from populations. To examine the relationship between the number of deleterious variants and recombination rate, we used a high-density genetic map from a soybean recombinant inbred line family (Lee et al. 2015). The soybean map was based on a subset of the SoySNP50K genotyping platform (Song et al. 2013). There was a weak but significant correlation between recombination rate and the proportion of nonsynonymous SNPs inferred to be deleterious ($r^2 = 0.007$, $p < 0.001$, Figures 3, S3). We did not examine this relationship in barley because the barley reference genome assembly (Mayer et al. 2012) contains limited physical distance information.

DISCUSSION

Questions regarding the prevalence of deleterious variants date back over half a century (Fisher 1930; Muller 1950). In finite populations, the segregation of deleterious variants can have a substantial impact on population mean fitness (Kimura et al. 1963). While it has been argued that the concept of a reduction of fitness relative to a hypothetical optimal genotype is irrelevant (Wallace 1970), mutation accumulation studies have shown that new mutations have a substantial effect on absolute fitness (Schultz et al. 1999; Shaw et al. 2002).

Our results demonstrate that a large number of putatively deleterious variants persist in individual cultivars in both barley and soybean. The approaches used in this study predict the probability that a given amino acid or nucleotide substitution disrupts

protein function. Mutations that alter phenotypes may be especially likely to annotate as deleterious, and we show that a high proportion of inferred causative mutations annotate as deleterious. It should be noted that variants identified as deleterious may affect a phenotype that is adaptive in only part of the species range or has a transient selective advantage – i.e., locally or temporally adaptive phenotypes. Our panel of causative variants consists primarily of SNPs that confer an agronomically important phenotype (Table S5). Agronomic phenotypes may be beneficial in wild populations, particularly biotic and abiotic stress tolerance or reproductive traits (Mercer et al. 2007), but are not expected to be either globally deleterious or globally beneficial. If the portion of the range in which the phenotype is adaptive is small or the selective advantage is transient, such variants will be kept at low frequencies and also be identified as deleterious. Just as few variants are expected to be globally advantageous, a portion of deleterious variation is likely to not be globally deleterious. Such variants could be either locally or temporally advantageous, with a fitness advantage under some circumstances contributing to their maintenance in populations (Tiffin and Ross-Ibarra 2014).

At the molecular level, variants occurring in minor transcripts of genes may exhibit conditional neutrality (Tiffin and Ross-Ibarra 2014), and $N_e s$ will be too low for purifying selection to act. Gan et al. (2011) identified many isoforms of genes among a diverse panel of *Arabidopsis thaliana* accessions, as well as compensatory mutations for a majority of frameshift mutations. Genetic variants that annotated as nonsynonymous or nonsense using the *A. thaliana* reference are frequently spliced out of the transcript such that the gene still produces a full-length and functional product. In a similar vein,

deleterious variants may have their fitness impacts offset by compensatory mutations. In a study of bacteriophage, approximately 70% of deleterious mutations were offset by compensatory mutations (Poon and Chao 2005). The occurrence of the bulk of putatively deleterious variants in the lowest frequency classes (Figure 1), and a higher level of observed heterozygosity for putatively deleterious variants (Figure S1) are both consistent with the action of purifying selection on variants with negative impacts on fitness. Putatively disease-causing variants in human populations have also been observed to occur at low frequencies and to occur over a more geographically restricted range (Marth et al. 2011).

Identifying variants with low minor allele frequency (MAF) is an inexorable part of studying variants with fitness impacts. This presents a problem, as rare variants are the most likely to be affected by false positive variant calls, as they are necessarily observed very few times in the sample. In an attempt to abate the problem of false positive variant calls, we took an iterative approach to variant calling, applying strict genotype quality, read depth, and observed heterozygosity filters to reduce raw variant calls to a high-confidence set of variants. While it is true that some of the variants in our high-confidence set are false positives, they do not dominate our dataset. This is evidenced in our site frequency spectra (Figure 1), which does not indicate a strong skewing of putatively neutral variants toward low frequency classes, a pattern expected of genotyping errors. Additionally, false positive variants are expected to occur randomly, which would lead to roughly equal numbers of first, second, and third position SNPs within codons. Our variant calls show a strong enrichment toward third positions in

codons (Figure S2), which are mostly synonymous positions, and are expected to be neutral. Deficiencies in first and second positions, which are mostly nonsynonymous sites under purifying selection, are indicative of our variant calls consisting mostly of true positive variants.

Comparison of Identification Methods

Each of the approaches used here to identify deleterious variants makes use of sequence constraint across a phylogenetic relationship. They differ in terms of the models used to assess the functional effect of a variant. SIFT uses a heuristic, which determines if a nonsynonymous variant alters a conserved site based on an alignment build from PSI-BLAST results (Ng 2003). Polyphen2 is similar, but additionally identifies potential disruptions in secondary or tertiary structure of the encoded protein (when this information is available) (Adzhubei et al. 2010), and is trained on known human disease-causing polymorphisms and neutral polymorphisms. Both of these approaches estimate codon conservation from a multiple sequence alignment, but do not use phylogenetic relationships in their predictions. PolyPhen2 identified the largest number of variants as deleterious, perhaps reflecting bias from the training dataset. Nonhuman systems may differ fundamentally as to which amino acid substitutions tend to have strong functional impact, which would reduce prediction accuracy in other species (Adzhubei et al. 2010). The LRT explicitly calculates the local synonymous substitution rate, and uses it to test whether an individual codon is under selective constraint or evolving neutrally (Chun and Fay 2009). It is a hypothesis-driven approach, and compares the likelihood of two evolutionary scenarios. Variants in selectively constrained codons are considered to be

deleterious.

Our results show that even though each prediction approach identifies a similar proportion of nonsynonymous SNPs as deleterious, the overlap between approaches is very small. Because each approach varies slightly in its prediction procedure and assumptions, the intersection of multiple approaches may provide more accurate predictions than any single prediction approach alone. At the genome-wide scale, this pattern is apparent in the frequency distribution of the variants that are identified as deleterious by all three approaches. They are enriched in the lowest frequency class, suggesting that they are under purifying selection (Figure 2). Further, known phenotype-altering SNPs are more likely to be predicted to be deleterious by all three approaches than those without known or measurable phenotypic impacts. This suggests that the intersection of prediction approaches tends to identify variants that are more likely to have fitness consequences, especially if the variant has a large effect on a phenotype. Identifying variants that are likely to have large effects on protein function and phenotype improves our ability to identify the nature of trait variation, especially if rare alleles of large effect are major contributors to complex traits (Thornton et al. 2013).

The SNPs predicted to be deleterious differ somewhat between prediction approaches. Even though SIFT and PolyPhen2 identify similar proportions of nonsynonymous SNPs as deleterious, they overlap at only approximately 50% of sites (Table 2). SNPs identified through at least two approaches seem more likely to be deleterious, based on lower average derived allele frequencies (Figure 1). Comparisons of the distribution of Grantham scores (Grantham 1974) show high similarity in the severity

of amino acid replacements that are predicted to be deleterious by each approach (Figure S6). The effects of reference bias are apparent in SIFT and PolyPhen2. In barley and soybean, the reference genotypes are 'Morex' and ‘Williams 82’ respectively. Even when polarized by ancestral and derived alleles, these genotypes show considerably fewer inferred deleterious variants (Table S6; Table S7).

The software package we have developed to implement the LRT is called BAD_Mutations (BLAST Aligned-Deleterious Mutations;). While BAD_Mutations is similar in approach to SIFT and PolyPhen2, it uses distinct data sources and models to predict whether or not a SNP is deleterious. SIFT and PolyPhen2 rely on BLAST searches against a general nucleotide sequence database, which results in high degree of variability in data quality from gene to gene (data not shown). BAD_Mutations, on the other hand, uses a set of assembled and annotated genome sequences available in the public domain in databases such as Phytozome (https://phytozome.jgi.doe.gov) and Ensembl Plants (http://plants.ensembl.org/). The use of a standard set of genome sequences helps to ensure consistent phylogenetic comparisons for each gene analyzed. It also uses a model that weights the conservation of the amino acid residue by the synonymous substitution rate of the gene under consideration (Chun and Fay 2009). BAD_Mutations is open source and freely available at https://github.com/MorrellLAB/BAD_Mutations.

Rise of Deleterious Variants Into Populations

The number of segregating deleterious variants in a species is very different from the number of *de novo* deleterious mutations in each generation, commonly identified as

U. *U* is the product of the per-base pair mutation rate, the genome size, and the fraction of the genome that is deleterious when mutated (Charlesworth 2012). In humans, *U* is estimated at ~2 new deleterious variants per genome per generation (Agrawal and Whitlock 2012). Estimates from *Arabidopsis thaliana* suggest that the genomic mutation rate for fitness-related traits is ~0.1-0.2 per generation (Shaw et al. 2002), and approximately half are estimated to be detrimental to fitness. Even though new mutations are constantly arising, the standing load of deleterious variation greatly exceeds the rate at which they arise (Charlesworth et al. 2004; Charlesworth 2012). However, our results show that ~40% of our inferred deleterious variants are private to individual cultivars, suggesting that they can be purged from breeding programs.

Once deleterious variants are present as segregating variation in the progenitors of crops, genetic bottlenecks associated with domestication (Eyre-Walker et al. 1998) may allow deleterious variants to drift to higher frequency (Robertson 1960). The selective sweeps associated with domestication and improvement (Wright et al. 2005) would decrease nucleotide diversity in affected genomic regions (Smith and Haigh 1974; Kaplan et al. 1989), and subsequently reduce the effective recombination rate (cf. O'Reilly 2008). The selective and demographic processes of domestication and improvement lead to three basic hypotheses about the distribution of deleterious variants in crop plants: *i*) the relative proportion of deleterious variants will be higher in domesticates than in wild relatives; *ii*) deleterious variants will be enriched near loci of agronomic importance that are subjected to strong selection during domestication and improvement; *iii*) the relative proportion of deleterious variants will be lower in elite

cultivars than landraces due to strong selection for yield (Gaut et al. 2015). Future studies of deleterious variants in crops and their wild relatives can address these hypotheses to understand the source of variation in modern cultivated material.

Deleterious Variants in Crop Breeding

The identification and targeted elimination of deleterious variants has been proposed as a potential means of improving plant fitness and crop yield (Morrell et al. 2011). Current plant breeding strategies using genomewide prediction rely on estimating genome-wide marker effects on quantitative traits of interest (Meuwissen et al. 2001). Genomewide prediction has been shown to be effective in both animals (Schaeffer 2006) and plants (Heffner et al. 2011; Jacobson et al. 2014), but these approaches rely on estimating marker contributions to a quantitative trait (i.e., a measured phenotypic effect). The genetic architecture of these traits suggests that our ability to quantify the effects of individual loci will reach practical limits before we can identify loci contributing to their variance (Rockman 2012). QTL mapping approaches to identifying favorable variants for agronomic traits will reach practical limits, even for variants of large effect (King et al. 2012). Many traits of agronomic interest, particularly yield in grain crops, are quantitative and have a complex genetic basis. As such, they are under the influence of environmental effects and many loci (Falconer and Mackay 1996). Current genomewide prediction and selection methodologies rely on estimating the combined effects of markers across the genome (Meuwissen et al. 2001), but this approach is limited by recombination rate and the ability to measure phenotypes of interest. The identification and purging of deleterious variants should provide a complementary approach to current

breeding methodologies, if bioinformatically-identified deleterious variants are truly deleterious (Morrell et al. 2011).

In the current study, we restricted our analyses to protein coding regions, but additional recent evidence suggests that deleterious variants can accumulate in conserved noncoding sequences, such as transcription factor binding sites (Arbiza et al. 2013). Additionally, insertion and deletion polymorphisms and larger structural variants were not considered in this study. Structural variants are abundant in crop plants, and may be involved with large phenotypic changes (Chia et al. 2012; Anderson et al. 2014). As such, analysis of nonsynonymous SNPs presents a lower bound on the estimates of the number of deleterious variants segregating in populations. Efforts to identify deleterious variants in noncoding sequence are limited by scant knowledge of functional constraints on noncoding genomic regions, and difficulty in aligning noncoding regions from all but the most closely related taxa (Doniger et al. 2008). Annotation of noncoding sequence will uncover additional deleterious variants, but the roughly one thousand putatively deleterious variants we identify per individual cultivar should provide ample targets for selection of recombinant progeny in a breeding program.

Species	Diff. From Ref	Noncoding	Syn.	Nonsyn.	Nonsense
Barley	162,954 (51,231.34)	115,456 (41,065.22)	15,591 (5,691.81)	12,351 (4,492.53)	77 (33.13)
Soybean	82,840 (56,780.03)	44,704 (29,477.65)	14,167 (8,161.21)	18,695 (11,289.72)	540 (345.05)

Table 1. Mean numbers of SNPs in various classes. Syn. = Synonymous; Nonsyn. = Nonsynonymous. Numbers are Mean (sd)

Species	SIFT	PPH	LRT	Intersect
Barley	3,400 (0.192)	3,295 (0.186)	3,221 (0.183)	1,006 (0.057)
Soybean	1,972 (0.064)	3,881 (0.126)	3,135 (0.101)	784 (0.025)

Table 2. Mean counts of nonsynonymous variants that are predicted to be deleterious by three prediction methods. Numbers in parentheses are proportions of all nonsynonymous variants in each sample that are predicted to be deleterious.

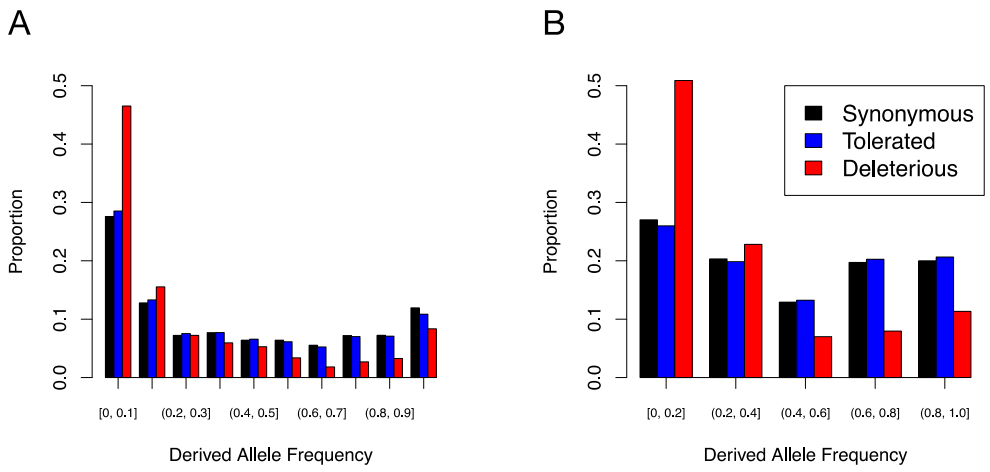

Figure 1: Derived allele (unfolded) frequency distributions for coding regions showing deleterious, tolerated, and synonymous SNPs for barley and soybean. Ancestral state was inferred as described in the methods. A variant was called "Deleterious" if it was nonsynonymous and predicted to be deleterious by SIFT, PolyPhen2, and the LRT. A) is based on thirteen domesticated barley accessions and two wild accessions while B) is based on seven cultivated soybean accessions and one wild accession.

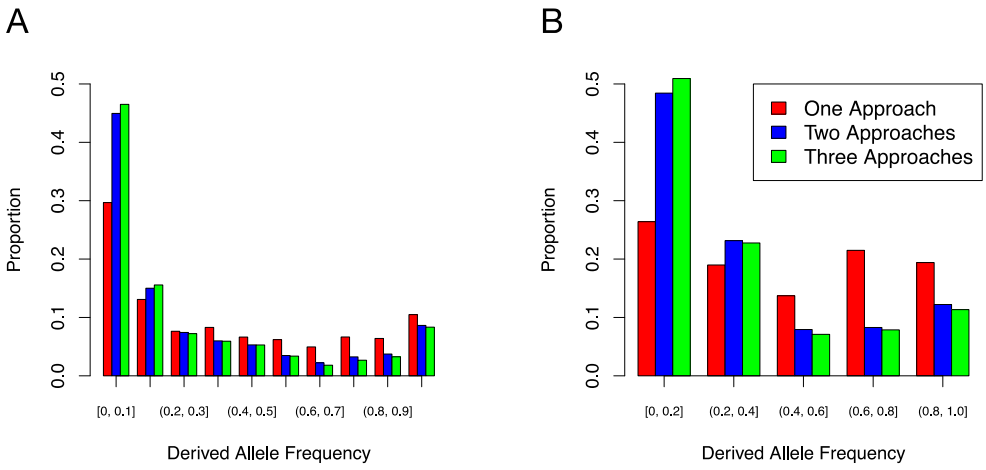

Figure 2: Derived allele (unfolded) frequency distributions for SNPs in A) barley and B) soybean predicted to be deleterious by one, two, or three prediction approaches. SNPs predicted by only one approach are not as strongly skewed toward rare variants, suggesting that the intersection of multiple prediction approaches gives the most reliable prediction of deleterious variants.

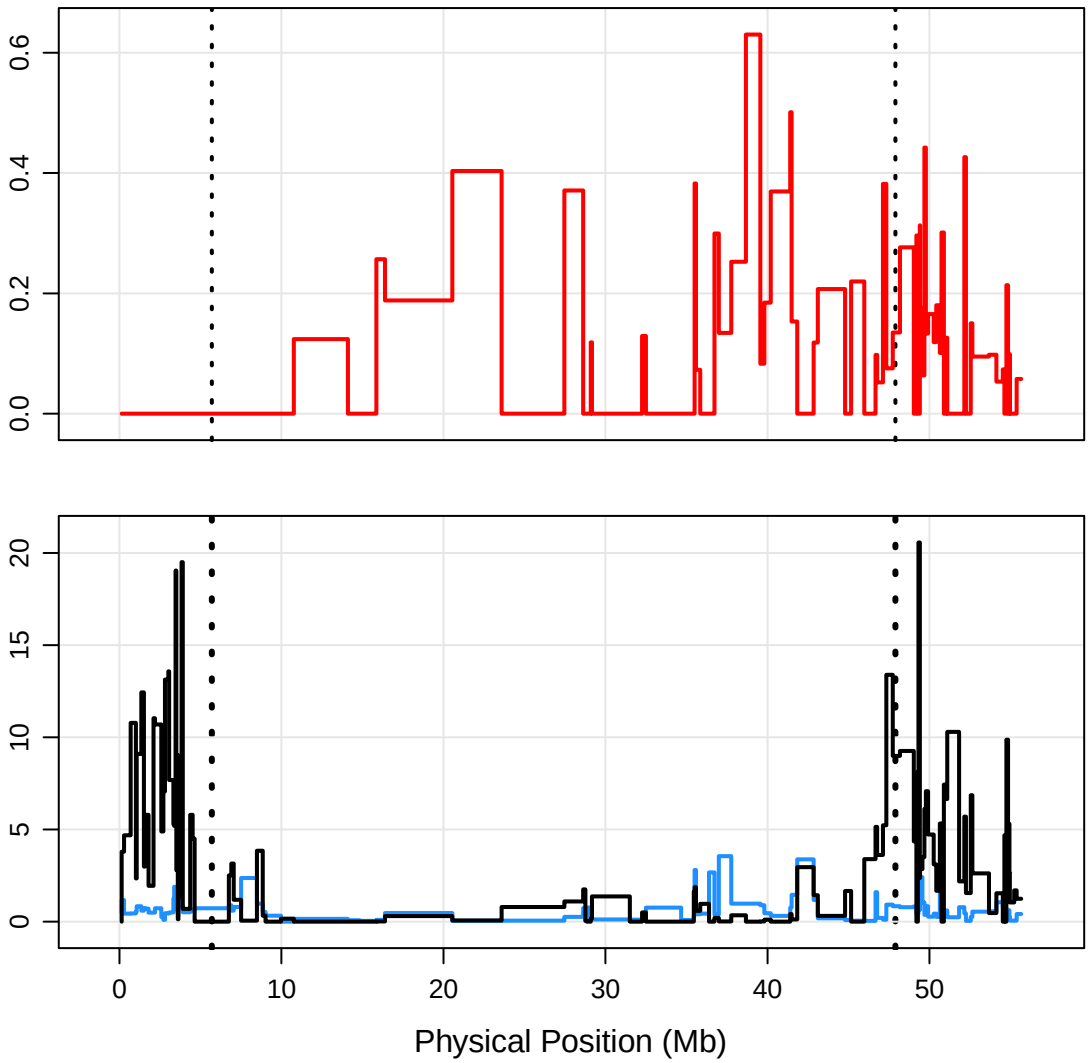

Figure 3: Comparison between recombination rate, coding sequence diversity, and proportion of nonsynonymous SNPs inferred to be deleterious in soybean on chromosome 1. The top panel shows the proportion of nonsynonymous SNPs that were inferred to be deleterious, in windows defined by genetic map distance (Lee et al. 2015). The bottom panel shows recombination rate in cM/Mb (black line) and average pairwise nucleotide sequence diversity per kilobase in coding sequence (blue line). Dashed vertical lines represent the boundaries of the annotated pericentromeric region, which has much lower recombination rates than the euochromatic regions.

Chapter 4: Estimating the relative contribution of deleterious and neutral SNPs to agronomic phenotypes

Targeted identification and purging of segregating deleterious alleles has been proposed as a novel approach to plant breeding. Recent advances in DNA resequencing technology and sequence constraint based approaches to predict the functional impact of a mutation now allow for identification of putatively deleterious SNPs on a genome-wide scale. Previous surveys of bioinformatically-identified deleterious SNPs show that individual crop genomes may carry hundreds of deleterious SNPs. However, the contributions of these SNPs to phenotypic variation have not been empirically evaluated. We use exome resequencing and SNP genotyping of three cycles of a spring six-row barley breeding population to compare the phenotypic contributions of putatively deleterious SNPs to those of neutral SNPs. In this population, grain yield remained stable while a proxy for disease severity decreased. We find that putatively deleterious SNPs explain more phenotypic variance for yield and disease severity on average than other functional classes of variants.

INTRODUCTION

An improved understanding of the genetic architecture of complex traits will assist in making gains from selection in plant and animal breeding. Complex traits are quantitative in nature, with contributions from both genetic and environmental factors (Falconer and Mackay 1996). One current source of debate is the relative frequency of genetic variants that contribute to quantitative trait variation. At mutation- drift

equilibrium, the majority of genetic variants segregating in a population are expected to be rare (Ewens 1972; Watterson 1975). If a genetic variant affects a phenotype, it is more likely to be subject to selection, with the strength of selection proportional to the magnitude of phenotypic impact. Since most new mutations with phenotypic impacts are expected to be deleterious (Thatcher et al. 1998; Sanjuán et al. 2004; Eyre-Walker et al. 2006) variants contributing to complex trait variation will likely be under purifying selection. Thus, a substantial portion of genetic variants that affect phenotypes may occur as "rare alleles of large effect" (RALE) (Thornton et al. 2013). In fact, association studies show evidence that rare alleles have larger estimated phenotypic effects than common alleles (Stanton-Geddes et al. 2013).

Practical limits on the statistical power to associate phenotypic variation with individual genetic variants cause association mapping studies to make a "common trait-common variant" (CTCV) assumption. Under this hypothesis, there is an implicit assumption that genetic variants contributing to phenotypic variation will occur in a sufficient proportion of the sample to exceed thresholds for statistical significance. While large effect variants may be readily detectable in this framework, they are unlikely to be representative of the majority of genetic variants that contribute to phenotypic variation (Rockman 2012). Larger sample sizes will be needed to observe variants of minor effect, or rare alleles that contribute to phenotypic variation. Ultimately, the samples or experimental populations necessary to identify variants across all frequency classes will exceed practical limits for many study systems (King et al. 2012).

Genome-wide prediction and selection methodology sidesteps the limitations of

association mapping by not assessing statistical significance of individual marker effects. The philosophy of genome-wide prediction is to use a panel with genotype and phenotype data to predict phenotype from only genotype in a population with high linkage disequilibrium (Meuwissen et al. 2001). Selection may then be performed on phenotypic values predicted based on genotypes. Selection based on genome-wide predictions has been demonstrated to work (Heffner et al. 2011; Jacobson et al. 2014), but it still relies on phenotypes being heritable. As such, predictions based on genotyped markers do not always match realized phenotypes (Lian et al. 2014).

Another potential problem with current implementations of genome-wide prediction and selection programs is the marker platforms used. SNPs on fixed genotyping chips have often been identified in a relatively shallow discovery panel and are unlikely to affect phenotypic variation in themselves. While the exact ascertainment scheme varies from marker panel to marker panel, these schemes generally involve SNPs that segregate at high minor allele frequency, and do not exhibit departures from neutral evolution (Close et al. 2009; Cavanagh et al. 2013). These properties make genotyped SNPs useful for distinguishing samples or populations, but less so for dissecting quantitative traits. Instead, identifying genome-wide variants with resequencing technology and using subsets of variants may improve accuracy. A study on the predictiveness of various variant classes for quantitative traits in *Drosophila melanogaster* found that prediction accuracy improved when differentially weighting variants that are more likely to affect phenotypes (Edwards et al. 2016). In particular, Edwards et al. (2016) demonstrate that when SNPs occur in genomic features that are annotated to be involved in known

biological processes, the predictive ability can increase by up to a factor of 2. Weighting partitions of variants that are likely to be enriched for functional variants, such as protein coding regions, may improve the performance of genome-wide prediction models.

The purpose of this study is to assess the degree to which variants identified via bioinformatics as likely to be deleterious contribute to variation in agronomic phenotypes. We used data collected from a barley breeding population developed at the University of Minnesota to address this goal. The major questions we sought to answer in this study were (1) How many putatively deleterious SNPs segregate in elite barley breeding material, and how many are private to individual lines or breeding programs? (2) What is the correlation between the number of putatively deleterious SNPs and yield? (3) What proportion of variance for yield do putatively deleterious SNPs explain, and is it different from other functional classes of markers?

MATERIALS AND METHODS

Population Design

Our experimental population consists of spring, six-row, malting barley adapted to the Upper Midwest region. A total of 21 breeding lines from three breeding programs (Busch Agricultural Resources, Inc., North Dakota State University, and University of Minnesota) were chosen as founders of the population, denoted as Cycle 0 (C0). Forty-five crosses were made with the parents. F1 progeny from each of the crosses were self-fertilized to the F_3 generation. Selections were made from 1,080 F_3 progeny. There were two selected pools - 50 lines chosen with the highest genomic estimated breeding value

(GEBV) and 50 lines chosen at random. The 100 chosen lines represent Cycle 1 (C1). The 50 lines selected by GEBV were then randomly intercrossed, and the resulting progeny were self-fertilized to the F_3 generation. The cycles of selection, intercrossing, and inbreeding were repeated two more times, resulting in three cycles of selection with genomic prediction. A schematic of the population design is shown in Figure S1.

Selections were carried out based on the predicted phenotypic values for grain yield and deoxynivalenol (DON) concentration, based on genomewide SNP markers. Lines were selected for increased yield and reduced DON concentration. GEBVs were estimated with ridge regression, as implemented in the 'rrBLUP' package for R.

Phenotypic Data Collection

Each GEBV and randomly selected line from each F_3 generation was evaluated in yield trials at five year-locations. Traits measured were grain yield in kilograms per hectare and DON concentration in parts per million. Phenotypic data were spatially adjusted with a moving average across the field plots. Best linear unbiased estimates (BLUEs) for yield and DON concentration were then produced for each line using the 'rrBLUP' package for R.

For yield trials in 2014, lines were evaluated at Crookston, MN; Morris, MN; and Saint Paul, MN. For 2015 yield trials, lines were evaluated at Crookston and Morris. Lines were grown in an augmented block design. The check varieties were 'Lacey' (96 replicates), 'Quest' (24 replicates), 'Stellar-ND' (20 replicates), and 'Tradition' (20 replicates).

For DON concentration trials, each chosen F_3 line was evaluated at five year-

locations. Similar to the yield trials, lines were grown in an augmented block design. DON concentration was evaluated at Crookston, MN in 2013, 2014, and 2015. DON concentration was evaluated at Saint Paul, MN in 2013 and 2014. Check varieties for DON trials were 'Quest' (123 replicates), 'ND20448' (26 replicates), 'Tradition' (25 replicates), and 'Lacey' (25 replicates).

Raw and adjusted phenotypic data, including planting locations in the field trials, are available at https://github.com/ MorrellLAB/Deleterious_GP.

Genotypic Data Collection

All F_3 progeny from each cycle of selection were genotyped with 384 SNP markers from the barley oligo pooled assay (BOPA) marker panel (Close et al. 2009). The markers were chosen to maximize the differences between the founders of the population. Genotypes were called using signal to noise ratios from the raw probe intensities, as implemented in ALCHEMY (Wright et al. 2010). ALCHEMY was chosen to call genotypes as it does not rely on clustering of samples to identify genotypic classes, and uses a prior on an inbreeding coefficient to model the number of expected heterozygous genotypes. The prior inbreeding coefficient was specified as 0.75 for each sample, which is the expected inbreeding coefficient after two generations of self-fertilization. Parental genotypes were imputed in progeny with BEAGLE 4.0 (Browning and Browning 2009). Imputed genotypes were set to missing if their genotype probability was less than 0.7.

DNA Extraction, Sequence Analysis, and Variant Calling

DNA was extracted from young leaf tissue from each of the 21 founder lines using a Plant DNAzol extraction reagent and protocol from Thermo Fisher Scientific

(Waltham, MA). Genomic DNA was captured with a 60 Mb liquid phase exome capture platform (Mascher et al. 2013). Eighteen of the samples were sequenced with 100bp paired end technology on an Illumina HiSeq2000, and three were sequenced with 125bp paired end technology on an Illumina HiSeq2500. The exomes were sequenced to a target depth of 30-fold coverage. Raw FASTQ files were cleaned of 3' sequencing adapter contamination with Scythe (https://github.com/vsbuffalo/scythe), using a prior on contamination rate of 0.05. Cleaned reads were then aligned to the Morex pseudo-molecule assembly with BWA-MEM (Li and Durbin 2009). Mismatch and alignment reporting parameters were tuned to allow for approximately three high-quality mismatches between the reads and the reference. This represents approximately the highest observed coding sequence diversity on barley (Morrell et al. 2006, 2014). The resulting BAM files were cleaned of unmapped reads, split alignments, and sorted with SAMtools version 1.3 (Li et al. 2009). Duplicate reads were removed with Picard version 2.0.1 (http://broadinstitute.github.io/picard/).

Alignment processing followed the Genome Analysis Toolkit (GATK) best practices workflow (McKenna et al. 2010; DePristo et al. 2011). Cleaned BAM alignments were realigned around putative insertion/deletion (indel) sites, as well as 100 previously identified indels in Sanger resequencing of 20 genes (Caldwell et al. 2006; Morrell et al. 2014). Individual sample genotype likelihoods were then calculated with the HaplotypeCaller, with a haploid model and "heterozygosity" value of 0.008 per base pair. This value is the mean estimate of coding nucleotide sequence diversity, based on previous Sanger resequencing experiments (Caldwell et al. 2006; Morrell et al. 2014).

SNP calls were made from the genotype likelihoods with GenotypeGVCFs. This first pass of SNP calls was then recalibrated, using Sanger resequencing variants (Caldwell et al. 2006; Morrell et al. 2014), BOPA SNP markers (Close et al. 2009), and ~5,000 SNP markers from the James Hutton Institute as a set of training variants. SNP calls passing filtration using previously identified SNPs as a training set were retained for analysis.

Scripts to perform adapter contamination removal, read mapping, alignment cleaning, and implementing the GATK best practices workflow are available at https://github.com/ MorrellLAB/Deleterious_GP.

Deleterious Predictions

Nonsynonymous SNPs were tested with three prediction approaches: PROVEAN (Choi et al. 2012), Polymorphism Phenotyping 2 (PPH2) (Adzhubei et al. 2013), and BAD_Mutations (Kono et al. 2016) which implements a likelihood ratio test for neutrality (Chun and Fay 2009). All three approaches use a form of sequence constraint test to predict whether a base substitution is likely to be deleterious. PROVEAN and PPH2 were run against the NCBI non-redundant protein sequence database, current as of 30 August, 2016. BAD_Mutations was run with a set of 42 publicly available Angiosperm genome sequences, hosted on Phytozome (https://phytozome.jgi.doe.gov) and Ensembl Plants (http://plants.ensembl.org/). Nonsynonymous SNPs were considered to be deleterious if they were identified as deleterious by all three approaches, or if they form an early stop codon (nonsense SNP). Previous results (Marth et al. 2011; Kono et al. 2016) show that SNPs that are predicted to be deleterious by multiple approaches are highly enriched for rare or private SNPs, showing that they are likely to be deleterious.

A single SNP was considered deleterious by PROVEAN if the substitution score was less than or equal to -4.1528, as determined by calculating 95% specificity from a set of known phenotype-altering SNPs in *Arabidopsis thaliana* (data not shown). A PPH2 prediction was considered as deleterious if it output a 'deleterious' call for a SNP. These programs use a heuristic approach for testing evolutionary constraint, as well as a training model for known human disease-causing polymorphisms. A SNP was considered deleterious by BAD_Mutations if the Bonferroni-corrected *p*-value was less than 0.05, with the number of affected codons representing the number of tests. Additional heuristics were applied to the BAD_Mutations results and sites with any of the following were considered as not deleterious: constraint greater than 1, more than ten gapped species in the alignment, a derived allele appearing in species other than the query species (barley). These heuristics help to ensure that predictions are made only on sites with substantial data to support estimation of evolutionary rates for the locus in question.

Proportion of Phenotypic Variance Explained

The proportion of phenotypic variance explained by various subsets of markers was estimated with a restricted maximum likelihood method implemented in GCTA (Yang et al. 2010, 2011). In order to compare the proportion of phenotypic variance explained by different subsets of SNPs, the number and frequency distribution of SNPs in each subset must be identical. Therefore, we took a resampling approach to compare the proportion of variance explained. Since putatively deleterious SNPs have the most restricted minor allele frequency distribution (Figure 2), we downsampled non-deleterious SNPs to match the minor allele frequency distribution of putatively deleterious SNPs. For each iteration,

the following steps were performed. First, a subset of SNPs was randomly sampled to match the frequency distribution of putatively deleterious SNPs in 1% frequency bins. Next, 500 SNPs were randomly sampled from the downsampled SNPs. A genetic relatedness matrix was then built from the 500 SNPs. Finally, the proportion of phenotypic variance explained by the random SNPs was estimated. The resampling procedure was performed 5,000 times for each partition of SNPs. To generate a null expectation for the proportion of phenotypic variance explained by 500 randomly chosen genomewide SNPs, the same resampling procedure was performed, with an additional step for randomizing the phenotypes with respect to the genotypes for each iteration.

Scripts for performing the resampling procedure to estimate the proportion of phenotypic variance explained are available at https://github.com/MorrellLAB/Deleterious_GP.

RESULTS

Plant Materials, Phenotypic Data, and Genotypic Data

The final dataset used for analysis consisted of 3,357 individuals, genotyped at 146,430 SNP markers. Of these, 140,279 were identified in exome capture sequencing of the parental varieties, and imputed into the population. The remaining 151 were BOPA SNPs that did not overlap with exome capture variants. From the panel of 3,357 individuals, a total of 317 of them have phenotypic data for yield and DON accumulation. It appears that the average DON concentration decreased over the course of the three cycles of selection, while yield remained stable (Figure 1).

Variant Calling and Deleterious SNP Identification

Of the 140,430 SNPs segregating in the parental lines, 34,791 were in protein coding sequence, and 17,027 affected codons, and 17,764 were synonymous. There were 242 nonsense variants identified. Among the nonsynonymous SNPs, 659 were inferred to be deleterious by three approaches. The total number of inferred deleterious SNPs segregating in this population is 901 (Table 1). Of all the SNPs segregating in the population, 20,875 had unambiguous inference of ancestral state based on alignments to *Brachypodium distachyon, Triticum urartu,* and *Aegilops tauschii*.

A derived site frequency spectrum showing synonymous, neutral nonsynonymous, and deleterious nonsynonymous SNPs is shown in Figure 2. Putatively deleterious SNPs tend to be private to individual lines: 399 (26.0%) of the putatively deleterious SNPs are found in only one parental line. Conversely 3,853 (18.5%) of genome wide SNPs with inferred ancestral state are private to an individual line. The number of putatively deleterious SNPs carried by an individual in the breeding program varied from 633 to 718 SNPs. The distribution of the number of putatively deleterious SNPs carried by each line follows an approximately normal distribution (Figure 3).

Putatively Deleterious SNPs and Phenotypic Variation

The number of putatively deleterious SNPs was not significantly correlated with grain yield (Figure S2, $p > 0.01$, $r=-0.097$). This is not surprising, as the number of putatively deleterious SNPs carried by individual lines does not change appreciably with cycles of selection (Figure 3), and grain yield remains stable with selection (Figure 1). On the other hand, the number of putatively deleterious SNPs was significantly and negatively correlated with DON concentration (Figure S3, $p < 0.01$, $r=-0.199$). This

implies that as a line accumulates more putatively deleterious SNPs, it decreases in severity of disease phenotypes. The relative frequencies of putatively deleterious and putatively neutral SNPs did not change with cycles of selection (Figure S4).

All partitions of genetic variants explained significant phenotypic variation for yield (Figure 4). Restricting the pool of genetic variants from all SNPs to only noncoding SNPs did not increase the proportion of phenotypic variance explained. Similarly, restricting from all coding SNPs to only nonsynonymous SNPs did not increase the proportion of phenotypic variance explained. However, putatively deleterious SNPs explained a larger proportion of yield variance than other partitions of SNPs (Figure 4). Similarly, for DON concentration, all partitions of SNPs explained significant variation for phenotypic variance (Figure 5). The partition of SNPs that explained the greatest proportion of phenotypic variance was noncoding SNPs.

DISCUSSION

In this barley breeding population, we identify 901 putatively deleterious SNPs segregating in protein coding regions. Many of the deleterious SNPs are private to an individual line, which provides circumstantial evidence that they are deleterious. The average minor allele frequency of putatively deleterious SNPs decreased over cycles of selection for increased agronomic performance, but the average number of putatively deleterious SNPs carried by individual lines did not change appreciably with selection. Despite this, putatively deleterious SNPs explained a greater proportion of phenotypic variance for yield than other partitions of SNPs on average.

An important limitation to these analyses is illustrated in the result that the number

of putatively deleterious SNPs carried by an individual line and its estimated DON concentration are negatively correlated. This implies that lines with less severe disease phenotypes have a higher number of putatively deleterious SNPs. While this appears to contradictory to the prediction of the effects of putatively deleterious SNPs on fitness, the fact that the phenotype is related to disease may explain the observation. Genetic loci that are involved with pathogen response may be under strong lineage-specific positive selection (Michelmore and Meyers 1998), and resistance-conferring SNPs in those genes may appear deleterious in our pipeline. Additionally, the variants identified as deleterious may affect phenotypes that are correlated with disease severity, such as flowering time and plant height.

A caveat of identifying deleterious variants through sequence constraint approaches is that mutations that alter a conserved amino acid residue are not necessarily deleterious. Algorithms that use only sequence constraint may identify variants that strongly affect protein function, but environmental conditions determine whether or not a variant is deleterious, neutral, or beneficial (Tiffin and Ross-Ibarra 2014). In some cases, a mutation that strongly impacts protein function is locally adaptive. This appears to be the case for many SNPs that are implicated with adaptation to agronomic environments (Kono et al. 2016). In many other cases, a variant at a conserved site may be deleterious, but only conditionally so. Selective coefficients against individual sequence variants fluctuate across time and geographic space (Tiffin and Ross-Ibarra 2014).

Deleterious variants segregate in cultivated material for several reasons. The first reason is simply that new mutations constantly arise, and finite N_e places limits on the

efficacy of selection against deleterious mutations (Crow and Kimura 1970). The other reasons relate to the demographic history of cultivated species and the action of linked selection. Most crop species experienced a bottleneck during domestication and subsequent bottlenecks during improvement. The reduced *N*e and strong inbreeding during these events reduces the efficacy of purifying selection, and deleterious variants may accumulate (Eyre-Walker et al. 1998; Wright et al. 2005). Additionally, strong selection for agronomic phenotypes can cause an increase in the frequency of deleterious variant that is associated with the selected allele. The selected allele may increase in frequency so quickly as to prevent recombination from dissociating the favorable variant from the deleterious variant (Smith and Haigh 1974; Kaplan et al. 1989). While selecting for high yield and uniformity in modern breeding programs may select against deleterious variants, there is potential for them to persist in regions of low recombination (McMullen et al. 2009).

One potential limitation of incorporating deleterious variants into genome-wide selection models is that most models only consider additive marker effects. This works well for SNPs on genotyping arrays, which are ascertained. That is, markers on genotyping arrays are biased toward common variants, which do not explain a large portion of heritable variation in themselves (Thornton et al. 2013). Deleterious variants, on the other hand, are expected to behave differently. On average, they are expected to be partially recessive (Simmons and Crow 1977; Shoemaker et al. 1996). They may also have compensatory variants that alleviate their effects, which is a form of epistatic interaction (Poon and Otto 2000; Poon and Chao 2005). These two forms of non-

additivity may complicate fitting models of phenotype as a function of genotype. In the case of barley, the issue is somewhat minor, as selections are performed on inbred lines, among which dominance variance is not expressed. Furthermore, previous efforts to identify epistatic effects on yield in barley revealed that genetic variation for yield is mostly additive (Xu and Jia 2007). However, it should be noted that the relative importance of non-additive effects on phenotypic variation depend on the study population, the phenotypes considered, and the environment in which the study is performed. While Xu and Jia (2007) do not find a large contribution from epistatic variance, it cannot be ruled out as contributing to variation among lines in this study. Caveats considered, our result of putatively deleterious SNPs contributing at least as much to phenotypic variation as putatively neutral SNPs suggests that "reverse genomic selection" against putatively deleterious SNPs may be a path forward in crop improvement.

Type of Variation	Count
SNP	146,279
Coding SNP	34,791
Nonsynonymous SNP	17,027
Early Stop Codon	242
Putatively Deleterious SNP	659

Table 1: Summary of SNPs identified in the exomes of the parental varieties.

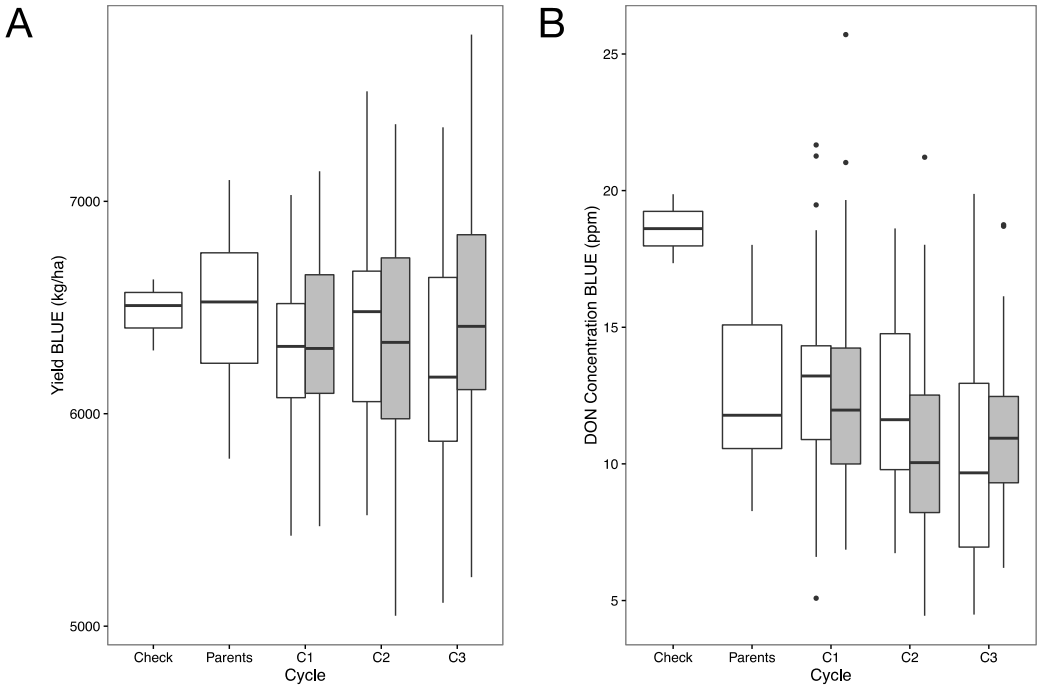

Figure 1: Plots of yield (A) and DON (B) data collected on this population. "Random" lines, check lines, and parental lines are shown in white, and "selected" lines are shown in grey. Selections were carried out based on genomic estimated breeding values. Data are best linear unbiased estimates (BLUEs) for individual lines based on yield and DON observations at five year-locations.

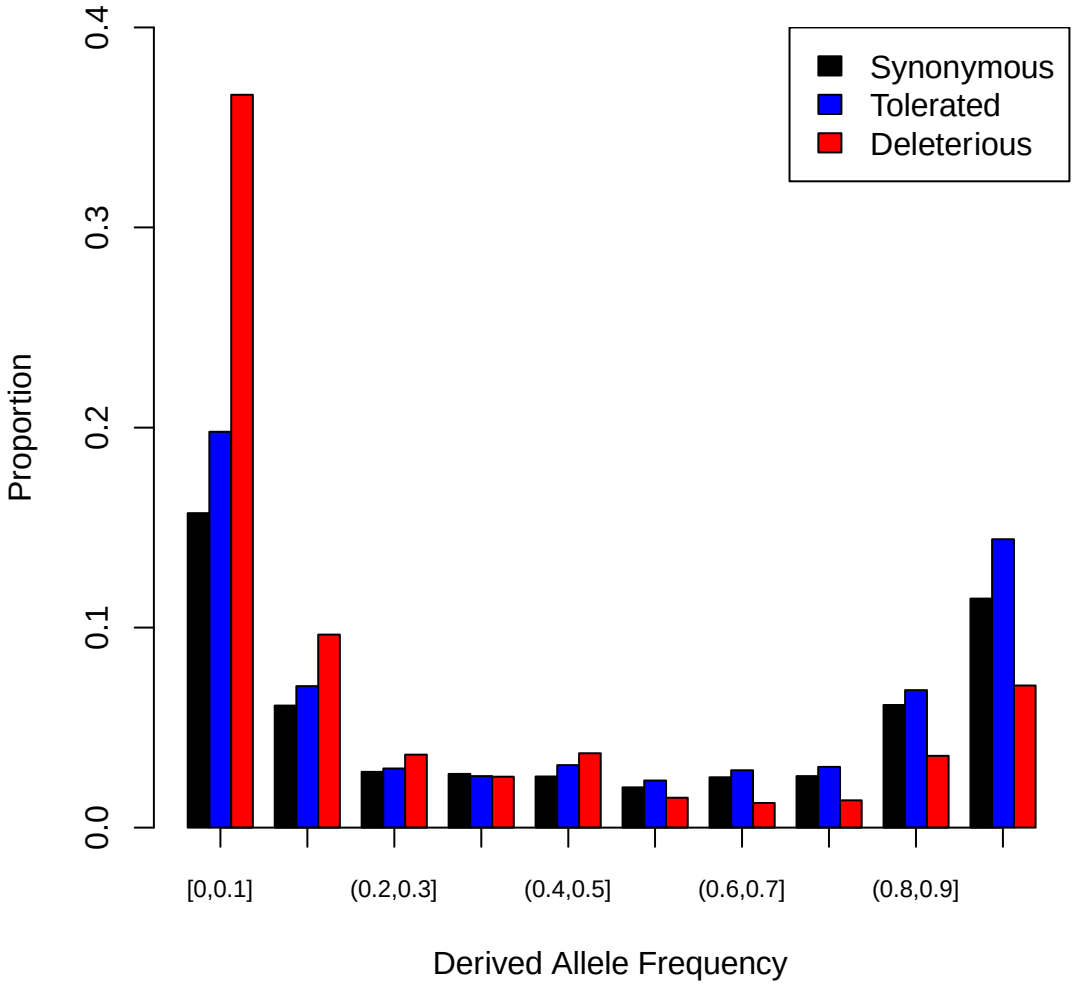

Figure 2: Derived site frequency spectrum showing synonymous (black), tolerated nonsynonymous (blue), and deleterious nonsynonymous (red) SNPs segregating in the parental lines of this population. Deleterious SNPs tend to segregate at lower derived allele frequency than neutral SNPs.

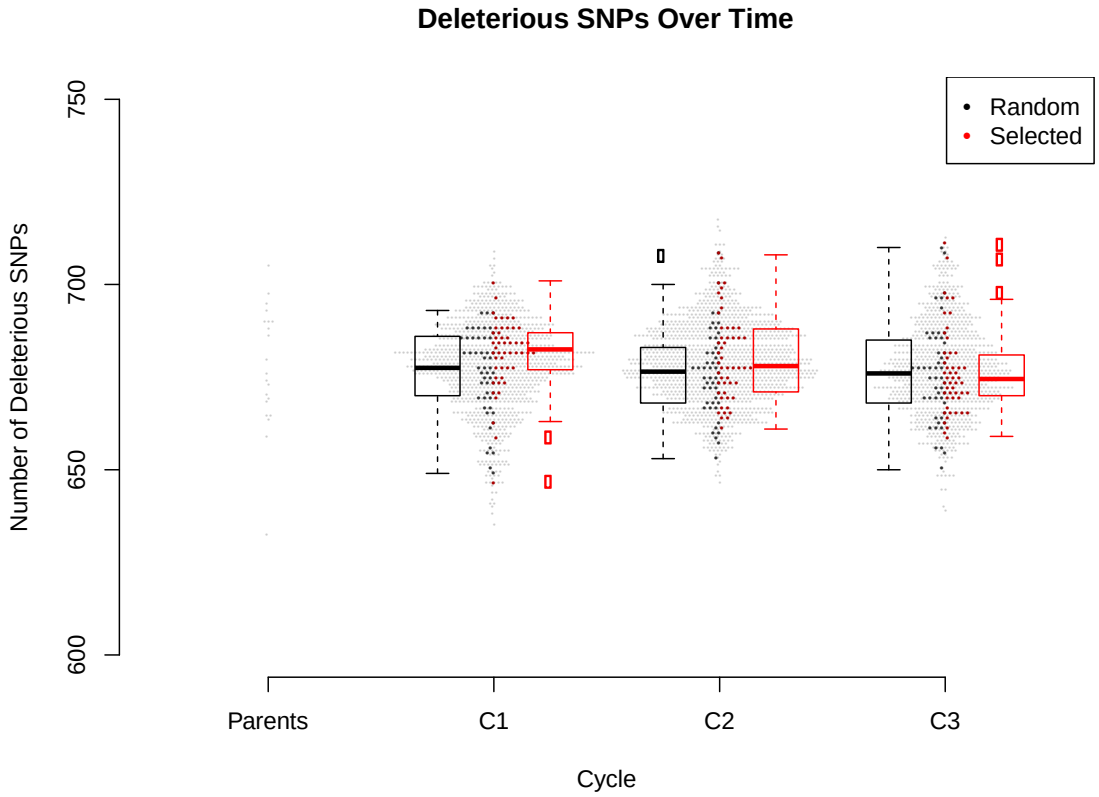

Figure 3: The number of deleterious SNPs carried by individual lines. Each grey point represents an individual in the population. Black points show lines that were randomly chosen from each cycle, and red points show lines that were selected based on genomic estimated breeding value for advancement of the breeding population. Boxplots show the distribution of the number of deleterious SNPs among random and selected lines. The number of deleterious SNPs does not change appreciably over cycles of selection.

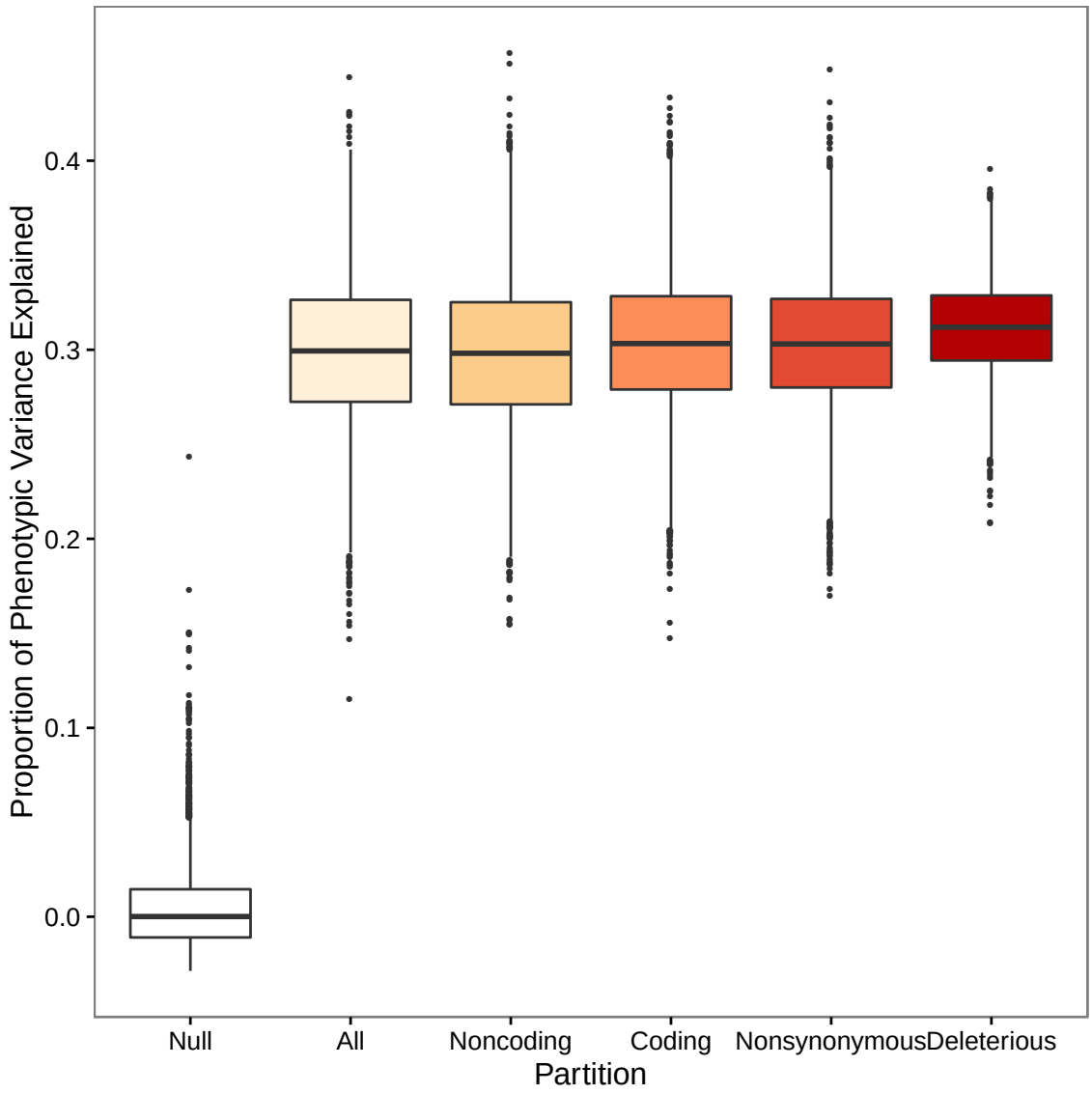

Figure 4: Boxplots showing the proportion of phenotypic variance for yield explained by partitions of the SNPs. “Null” is the proportion explained by randomized phenotypes. "All" refers to genome wide SNPs, "Coding" refers to SNPs in protein coding regions, "Noncoding" refers to SNPs in regions that do not code for proteins, and "Nonsynonymous" refers to SNPs that alter a protein amino acid sequence. Putatively deleterious SNPs explain more phenotypic variation on average than other partitions.

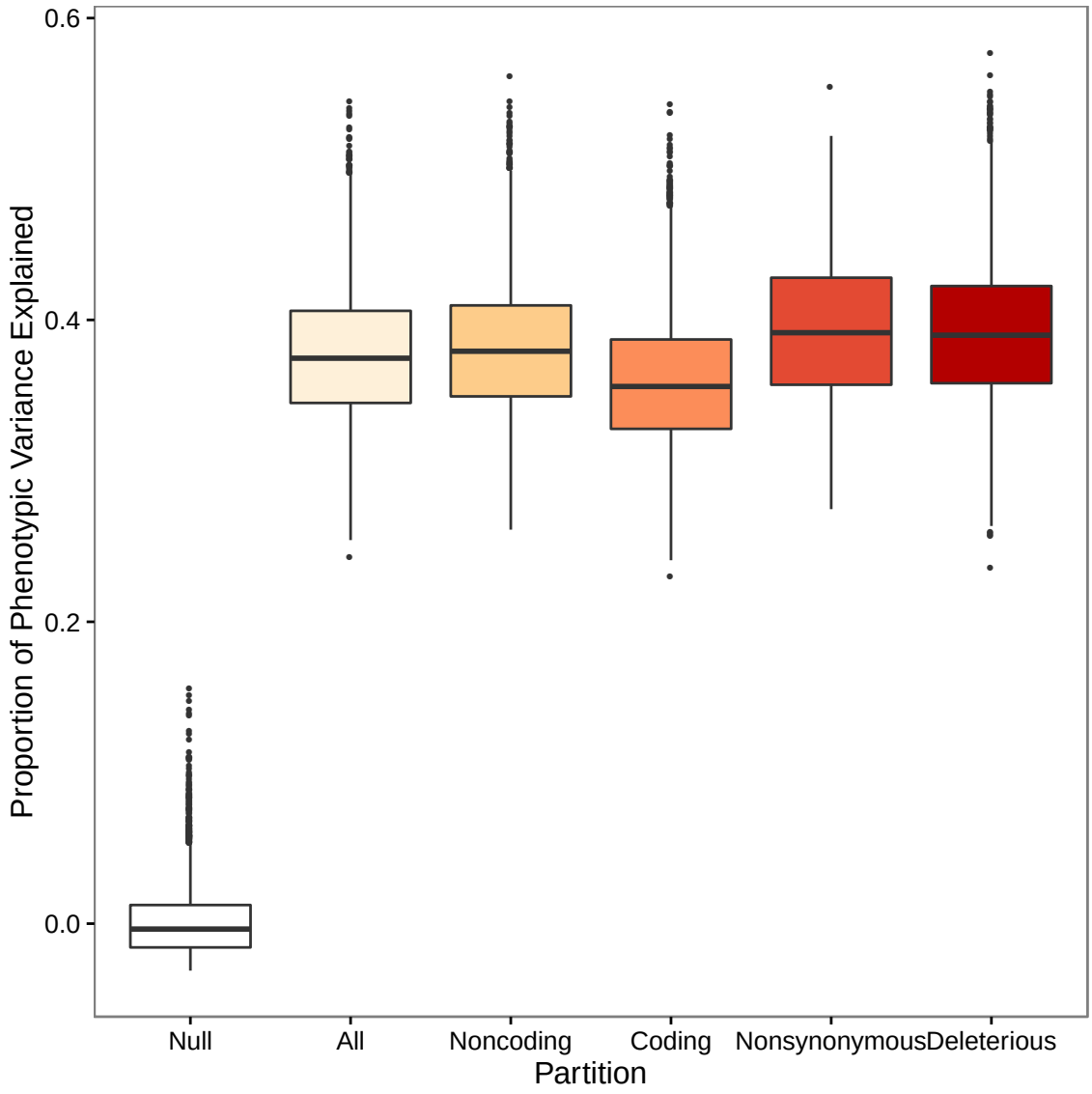

Figure 5: Boxplots showing the proportion of phenotypic variance for DON concentration explained by partitions of the SNPs. The partitions are the same as in Figure 4. Nonsynonymous and putatively deleterious SNPs explain more variation for DON concentration than other partitions of SNPs.

Bibliography

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature*. 467:1061-1073.

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. Chapter 7:Unit7.20.

Agrawal AF, Whitlock MC. 2012. Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. *Annu Rev Ecol Evol Syst*. 43:115-135.

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. Mckusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res*. 37:D793-D796.

Anderson JE, Kantar MB, Kono TY, Fu F, Stec AO, Song Q, Cregan PB, Specht JE, Diers BW, Cannon SB et al. 2014. A roadmap for functional structural variants in the soybean genome. *G3 (Bethesda)*. 4:1307-1318.

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res*. 21:610-617.

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A. 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet*. 45:723-729.

Arnheim N, Calabrese P, Nordborg M. 2003. Hot and cold spots of recombination in the

human genome: The reason we should find them and how this can be achieved. *Am J Hum Genet*. 73:5-16.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*. 3:e3376.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.

Blake VC, Kling JG, Hayes PM, Jannink J-L, Jillella SR, Lee J, Matthews DE, Chao S, Close TJ, Muehlbauer GJ et al. 2012. The *Hordeum* toolbox: The barley Coordinated Agricultural Project genotype and phenotype resource. *Plant Genome*. 5:81-91.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*. 84:210-223.

Caldwell KS, Russell J, Langridge P, Powell W. 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics*. 172:557-567.

Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining regions of the *drosophila melanogaster* genome. *Genome Biol Evol*. 4:278-288.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between

recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Mol Biol Evol*. 31:1010-1028.

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A et al. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc Natl Acad Sci U. S. A.* 110:8057-8062.

Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R, Gornicki P. 2008. Acc homoeoloci and the evolution of wheat genomes. *Proc Natl Acad of Sci U. S. A.* 105:9691-9696.

Charlesworth B. 2012. The effects of deleterious mutations on evolution at linked sites. *Genetics*. 190:5-22.

Charlesworth B, Borthwick H, Bartolomé C, Pignatelli P. 2004. Estimates of the genomic mutation rate for detrimental alleles in *Drosophila melanogaster*. *Genetics*. 167:815-826.

Charlesworth B, Charlesworth D. 1999. The genetic basis of inbreeding depression. *Genet Res*. 74:329-340.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*. 44:803-807.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. *Genome Res*. 19:1553-1561.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden

DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SNPEff: SNPs in the genome of *Drosophila melanogaster* strain *w*1118; *iso*-2; *iso*-3. *Fly*. 6:80-92.

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S et al. 2009. Development and implementation of high-throughput snp genotyping in barley. *BMC Genomics*. 10:582.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B et al. 2009. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25:1422-1423.

Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, Shaw P, Bayer M, Thomas W, Marshall D et al. 2012. Natural variation in a homolog of *antirrhinum* centroradialis contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet*. 44:1388-1392.

Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD. 2010. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat*. 31:631-655.

Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York, Evanston and London: Harper & Row, Publishers.

Crow JF. 1958. Some possibilities for measuring selection intensities in man. *Human Biology*.

Crow JF. 1963. 2. The concept of genetic load: A reply. *Am J Hum Genet*.

Cruz F, Vilà C, Webster MT. 2008. The legacy of domestication: Accumulation of deleterious mutations in the dog genome. *Mol Biol Evol*. 25:2331-2336.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nat Rev Genet*. 14:262-274.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*. 12:499-510.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 43:491-498.

Dillon SL, Shapter FM, Henry RJ, Cordeiro G, Izquierdo L, Lee LS. 2007. Domestication to crop improvement: Genetic resources for *Sorghum* and *Saccharum* (Andropogoneae). *Annals of Botany*. 100:975-989.

Dobzhansky T. 1937. Genetics and the origin of species. New York: Columbia University Press.

Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, Fay JC. 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet*. 4:e1000183.

Edwards SM, Sørensen IF, Sarup P, Mackay TF, Sørensen P. 2016. Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics*. 203:1871-1883.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*. 6:e19379.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol*. 3:87-112.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 8:610-618.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*. 173:891-900.

Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. 1998. Investigation of the bottleneck leading to the domestication of maize. *Proc Natl Acad Sci U. S. A.* 95:4441-4446.

Falconer DS, Mackay TFC. 1996. Introduction to quantitative genetics. London: Longman.

Fay JC, Wyckoff GJ, Wu C-I. 2001. Positive and negative selection on the human genome. *Genetics*. 158:1227-1234.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics*. 78:737-756.

Fisher RA. 1930. The genetical theory of natural selection. London: Oxford University Press.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*. 477:419-423.

Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M. 2015. An estimate of the average number of recessive lethal mutations carried by humans. *Genetics*. 199:1243-1254.

Gaut BS, Díez CM, Morrell PL. 2015. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet*. 31:709-719.

Glémin S. 2003. How are deleterious muations purged? Drift versus nonrandom mating. *Evolution*. 57:2678-2687.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science*. 185:862-864.

Graur D, Li W-H. 2000. Fundamentals of molecular evolution. Sunderland, MA: Sinauer Associates.

Günther T, Schmid KJ. 2010. Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. *Theor Appl Genet*. 121:157-168.

Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science*. 51:2597.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. *Nat Rev Genet*. 16:333-343.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K et al. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*. 26:121-135.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC,

Cregan PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci U S A*. 103:16666-16671.

Jacobson A, Lian L, Zhong S, Bernardo R. 2014. General combining ability model for genomewide selection in a biparental cross. *Crop Science*. 54:895.

Kane NC, Gill N, King MG, Bowers JE, Berges H, Gouzy J, Bachlava E, Langlade NB, Lai Z, Stewart M et al. 2011. Progress towards a reference genome for sunflower. *Botany*. 89:429-437.

Kaplan NL, Hudson RR, Langley CH. 1989. The" hitchhiking effect" revisited. *Genetics*. 123:887-899.

Kim MY, Lee S, Van K, Kim TH, Jeong SC, Choi IY, Kim DS, Lee YS, Park D, Ma J et al. 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc Natl Acad Sci U S A*. 107:22032-22037.

Kimura M, Maruyama T, Crow JF. 1963. The mutation load in small populations. *Genetics*. 48:1303.

King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD, Macdonald SJ. 2012. Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. *Genome Res*. 22:1558-1566.

Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. 2016. The role of deleterious substitutions in crop genomes. *Mol Biol Evol*. 33:2307-2317.

Lande R, Schemske DW. 1985. The evolution of self-fertilization and inbreeding

depression in plants. I. Genetic models. *Evolution*. 24-40.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 9:357-359.

Lee S, Freewalt KR, McHale LK, Song Q, Jun T-H, Michel AP, Dorrance AE, Mian MAR. 2015. A high-resolution genetic linkage map of soybean based on 357 recombinant inbred lines genotyped with BARCSoySNP6K. *Molecular Breeding*. 35:1-7.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, et al. 2009. The sequence alignment/map format and SAMTools. *Bioinformatics*. 25:2078-2079.

Lian L, Jacobson A, Zhong S, Bernardo R. 2014. Genomewide prediction accuracy within 969 maize biparental populations. *Crop Science*. 54:1514.

Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X et al. 2011. The functional spectrum of low-frequency coding variation. *Genome Biol*. 12:R84.

Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, Pfeifer M, D'Ascenzo M et al. 2013. Barley whole exome capture: A tool for genomic research in the genus *Hordeum* and beyond. *Plant J*. 76:494-505.

Matsumoto T, Tanaka T, Sakai H, Amano N, Kanamori H, Kurita K, Kikuta A, Kamiya

K, Yamamoto M, Ikawa H et al. 2011. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol*. 156:20-28.

Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ et al. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature*. 491:711-716.

Mayer KFX, Martis M, Hedley PE, Šimková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H et al. 2011. Unlocking the barley genome by chromosomal and comparative genomics. *The Plant Cell*. 23:1249-1263.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20:1297-1303.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C et al. 2009. Genetic properties of the maize Nested Association Mapping population. *Science*. 325:737-740.

McQuilton P, St. P, Susan E., Thurmond J, Gelbart W, Brown N, Kaufman T, Matthews K, Werner-Washburne M, Cripps R, Crosby L et al. 2012. FlyBase 101 - the basics of navigating FlyBase. *Nucleic Acids Res*. 40:D706-D714.

Mercer KL, Andow DA, Wyse DL, Shaw RG. 2007. Stress and domestication traits increase the relative fitness of crop-wild hybrids in sunflower. *Ecol Lett*. 10:383-393.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using

genome-wide dense marker maps. *Genetics*. 157:1819-1829.

Mezmouk S, Ross-Ibarra J. 2014. The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)*. 4:163-171.

Mirab S, Nguyen N, Warnow T. 2014. PASTA: ultra-large multiple sequence alignment. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B et al., editors. Cham: Springer International Publishing. p. 177-191.

Morrell PL, Buckler ES, Ross-Ibarra J. 2011. Crop genomics: Advances and applications. *Nat Rev Genet*. 13:85-96.

Morrell PL, Clegg MT. 2007. Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. *Proc Natl Acad Sci U. S. A.* 104:3289-3294.

Morrell PL, Gonzales AM, Meyer KK, Clegg MT. 2014. Resequencing data indicate a modest effect of domestication on diversity in barley: A cultigen with multiple origins. *J Hered*. 105:253-264.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2006. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. *Genetics*. 173:1705-1723.

Mukai T. 1964. The genetic structure of natural populations of drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics*. 50:1-19.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res-Fund Mol M*. 1:2-9.

Muller HJ. 1950. Our load of mutations. *Am J Hum Genet*. 2:111.

Ng PC. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 31:3812-3814.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics*. 154:923-929.

O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the ped/pop method for detecting selection. *Genome Res*. 18:1304-1313.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*. 327:92-94.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*. 7:e32253.

Poon A, Chao L. 2005. The rate of compensatory mutation in the DNA bacteriophage phix174. *Genetics*. 170:989-999.

Poon A, Otto SP. 2000. Compensating for our load of mutations: Freezing the meltdown of small populations. *Evolution*. 54:1467-1479.

Renaut S, Rieseberg LH. 2015. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other Compositae crops. *Mol Biol Evol*. 32:2273-2283.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open

Software Suite. *Trends in Genetics*. 16:276-277.

Robertson A. 1960. A theory of limits in artificial selection. *P Roy Soc Lond B Bio*. 153:234-249.

Rockman MV. 2012. The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*. 66:1-17.

Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, Li C, Li Y, Buckler ES. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc Natl Acad Sci U. S. A.* 112:3823-3828.

Rosenberg NA. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*. 57:1465-1477.

Sanghvi LD. 1963. 1. The concept of genetic load: A critique. *Am J Hum Genet.*

Sanjuán R, Moya A, Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A*. 101:8396-8401.

Sato K, Shin-I T, Seki M, Shinozaki K, Yoshida H, Takeda K, Yamazaki Y, Conte M, Kohara Y. 2009. Development of 5006 full-length cDNAs in barley: A tool for accessing cereal genomics resources. *DNA Research*. 16:81-89.

Schaeffer LR. 2006. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet*. 123:218-223.

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

Nature. 463:178-183.

Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, Albrechtsen A, Dupanloup I, Foucal A, Petersen B et al. 2014. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc Natl Acad Sci U S A*. 111:E5661-9.

Schultz ST, Lynch M, Willis JH. 1999. Spontaneous deleterious mutation in *Arabidopsis thaliana*. *Proc Natl Acad Sci U. S. A*. 96:11393-11398.

Shaw FH, Geyer CJ, Shaw RG. 2002. A comprehensive model of mutations affecting fitness and inferences for *Arabidopsis thaliana*. *Evolution*. 56:453-463.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res*. 29:308-311.

Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet*. 14:450-456.

Simmons MJ, Crow JF. 1977. Mutations affecting fitness in *Drosophila* populations. *Annu Rev Genet*. 11:49-78.

Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. *Nat Genet*. 46:220-224.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res*. 23:23.

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB. 2013. Development and evaluation of SoySNP50k, a high-density genotyping array for soybean. *PLoS One*. 8:e54985.

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, Bharti AK, Farmer AD, Zhou P, Denny R et al. 2013. Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. *PLoS One*. 8:e65688.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Mol Biol Evol*. 28:1569-1580.

Thatcher JW, Shaw JM, Dickinson WJ. 1998. Marginal fitness contributions of nonessential genes in yeast. *Proc Natl Acad Sci U. S. A*. 95:253-257.

Thornton KR, Foran AJ, Long AD. 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.

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

Wallace B. 1970. Genetic load, its biological and conceptual aspects. Prentice-Hall.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7:256-276.

Weigel D, Mott R. 2009. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol*. 10:107.

Wright MH, Tung CW, Zhao K, Reynolds A, McCouch SR, Bustamante CD. 2010. ALCHEMY: A reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. *Bioinformatics*. 26:2952-2960.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. *Science*. 308:1310-1314.

Xu S, Jia Z. 2007. Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics*. 175:1955-1963.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 42:565-569.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet*. 88:76-82.

Zhang Z, Schwartz S, Wagner L, Miller W. 2004. A greedy algorithm for aligning DNA sequences. *J Comput Biol*. 7:203-214.

Appendix 1: Chapter 2 Supplementary Material

SUPPLEMENTARY TEXT

SNPMeta accepts SNP contextual sequence in FASTA format, either as a single file or a set of files. Sequences are used for BLAST queries against the NCBI 'nucleotide' (nt) database. Options passed on the command line can be used to limit the number or scope of BLAST results (hits). Results are returned as an XML BLAST report, and parsed using the Bio.SeqIO module in BioPython (Cock et al., 2009). GenBank accession numbers, hit positions and sequence orientations are extracted from the BLAST report.

GenBank records for each SNP contextual sequence are retrieved from NCBI servers. A maximum of 5 kb per record is downloaded, 2.5 kb on either side of the hit position. The amount of sequence to download was determined by trial and error: it allows for identification of nearby genes, and avoids overloading NCBI's servers. SNPMeta then iterates through the list of records, saving both the first record (best match to query sequence) with an annotated coding sequence (CDS), and the first record with an annotated gene. Given that a GenBank record with a CDS annotation is found, SNPMeta extracts the positions of the sequence that the CDS covers. The record containing an annotated gene is used to identify potential homologues of the protein used for annotation. If there are no hits with an annotated CDS, SNPMeta prints a 'No Annotations' message and continues to query additional SNP contextual sequences.

Given that the query sequence identifies an annotated sequence in GenBank, the query and GenBank hit are locally aligned using the 'needle' program in the EMBOSS (Rice et al., 2000) suite. Then, using the contextual sequence length provided by the user, SNPMeta calculates the position of the SNP in the alignment. The positions of the CDS annotation in the GenBank record are used to determine if the SNP occurs in CDS. If the SNP is outside a coding region, SNPMeta calculates the distance from the SNP to the nearest identified segment of CDS.

If the SNP is in a coding region, SNPMeta uses the position of the SNP in the CDS to identify the codon that contains the SNP. The zero-based position of the SNP in the CDS is divided by three, and the remainder (zero, one, or two) indicates the position within the codon at which the SNP occurs (first, second or third, respectively). The codon configuration is determined by the following formula: the start of the codon is the position of the SNP minus the remainder of the above division, and the end of the codon is the position of the SNP plus two, minus the remainder of the division. The ambiguous nucleotide is then inserted into the resulting codon. The polymorphic codons are then translated using the translation table provided in the GenBank record. If no translation table is defined, it defaults to the standard codon translation table. Translations are used to determine the nature of the mutation – synonymous or nonsynonymous.

Several issues can arise in the annotation pipeline, particularly in the alignment step. SNPMeta is written to gracefully handle pathological alignments. Alignment errors most frequently involve gaps. For instance, it is possible that the SNP contextual

sequence aligns to the GenBank sequence such that the SNP does not occur within the GenBank sequence itself. Additionally, the query SNP could align over a gap in the GenBank sequence, or a large break in the CDS (for example, if the SNP is not in an exon). One commonly encountered problem is comparison of genomic sequence to cDNA, either because the SNP contextual sequence was designed such that it spans an intron-exon boundary, or because a SNP contextual sequence that includes an intron-exon boundary only matches cDNA records in GenBank. Both of these circumstances produce poor alignment scores, which causes SNPMeta to print a warning into the 'Notes' field. Lastly, the state of the SNP in the GenBank sequence might not be represented by the IUPAC ambiguity code. In all these cases, there is no annotation information produced. Instead, a message is printed, indicating further investigation by the researcher may be necessary.

SUPPLEMENTARY FIGURES

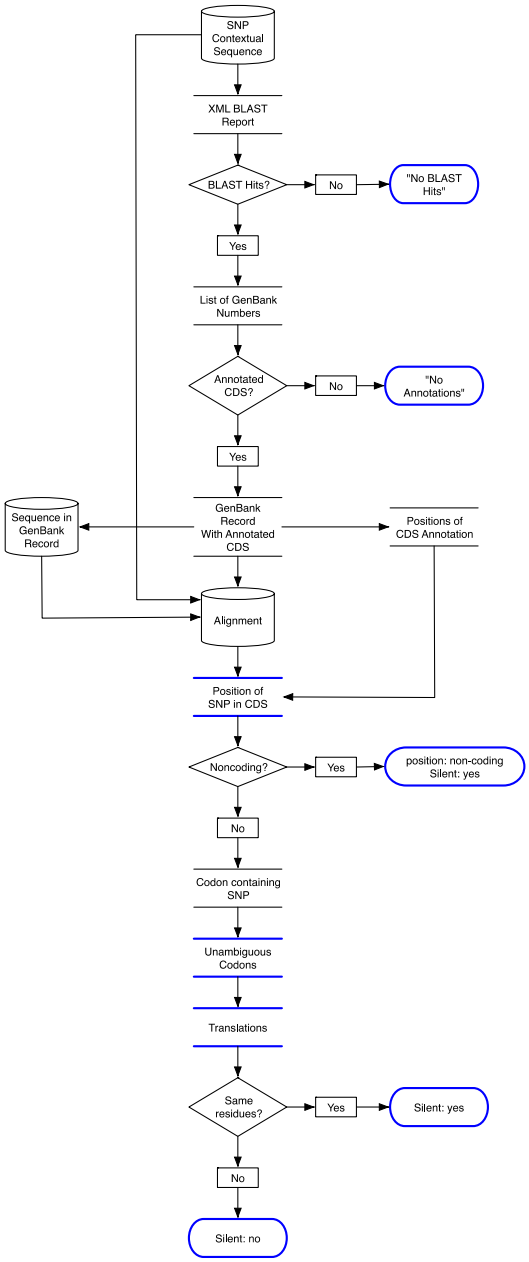

Figure. S1: Detailed view of the SNPMeta annotation workflow. Items in blue represent information that is included in the annotation output.

Appendix 2: Chapter 3 Supplementary Material

SUPPLEMENTARY TABLES

Species	Sample	SRA Accession(s)	Target	Final Coverage	Technology
Barley	Barke	ERX245921-ERX245931	Exome	28	PE 101bp
Barley	Bonus	ERX245948	Exome	38	PE 101bp
Barley	Borwina	ERX245949	Exome	59	PE 101bp
Barley	Bowman	ERX245945	Exome	29	PE 101bp
Barley	Foma	ERX245950	Exome	49	PE 101bp
Barley	Gull	ERX245951	Exome	38	PE 101bp
Barley	Harrington	ERX245955	Exome	31	PE 101bp
Barley	Haruna Nijo	ERX245953	Exome	38	PE 101bp
Barley	Igri	ERX245954	Exome	37	PE 101bp
Barley	Kindred	ERX245952	Exome	42	PE 101bp
Barley	Morex	ERX245932-ERX245942	Exome	35	PE 101bp
Barley	Steptoe	ERX245946, ERX245947	Exome	33	PE 101bp
Barley	Vogelsanger Gold	ERX245956	Exome	57	PE 101bp
Wild Barley	FT11	ERX245963	Exome	45	PE 101bp
Wild Barley	OUH602	ERX245964	Exome	33	PE 101bp
Soybean	Archer	SRX1250057	Whole Genome	37	PE 151bp
Soybean	Minsoy	SRX1541707	Whole Genome	37	PE 151bp
Soybean	Noir1	SRX1250058	Whole Genome	33	PE 151bp
Soybean	Wm82 ISU	SRX552896	Whole Genome	14	PE 101bp
Soybean	Williams	SRX1541708	Whole Genome	28	PE 101bp
Soybean	M92-220	SRX467183	Whole Genome	56	PE 101bp
Soybean	IA3023	SRX551636	Whole Genome	47	PE 151bp
Glycine soja	—	SRA009252	Whole Genome	52	SE 36bp PE 76bp

Table S1: Accessions used in this study. The final coverage reported is the average depth over the targeted region.

Species	Common Name	Assembly Version	Annotation Version	Downloaded From
Aegilops tauschii	Goatgrass	ASM34733v1	1	Ensembl Plants
Aquilegia coerulea	Columbine	1.1	1.1	DOE-JGI; Phytozome 10
Arabidopsis lyrata	Lyrate rockcress	1	1	DOE-JGI; Phytozome 10
Arabidopsis thaliana	Thale cress	TAIR10	TAIR10	DOE-JGI; Phytozome 10
Boechera stricta	Drummond's rockcress	1.2	1.2	DOE-JGI; Phytozome 10
Brachypodium distachyon	Purple false brome	2.1	2.1	DOE-JGI; Phytozome 10
Brassica oleracea	Cabbage	2.1	2.1	Ensembl Plants
Brassica rapa	Turnip mustard	FPsc 1.3	1	DOE-JGI; Phytozome 10
Capsella grandiflora		1.1	1.1	DOE-JGI; Phytozome 10
Capsella rubella	Red shepherd's purse	1	1	DOE-JGI; Phytozome 10
Carica papaya	Papaya	ASGPBv0.4	ASGPBv0.4	DOE-JGI; Phytozome 10
Citrus clementina	Clementine	1	clementine1.0	DOE-JGI; Phytozome 10
Citrus sinensis	Sweet orange	1	orange1.1	DOE-JGI; Phytozome 10
Cucumis sativus	Cucumber	1	1	DOE-JGI; Phytozome 10
Eucalyptus grandis	Eucalyptus	2	2	DOE-JGI; Phytozome 10
Eutrema salsugineum	Salt cress	1	1	DOE-JGI; Phytozome 10
Fragaria vesca	Strawberry	1.1	1.1	DOE-JGI; Phytozome 10
Glycine max	Soybean	a2	a2.v1	DOE-JGI; Phytozome 10
Gossypium raimondii	Cotton	2.1	2.1	DOE-JGI; Phytozome 10

Hordeum vulgare	Barley	082214v1	1	Ensembl Plants
Leersia perrieri	Cutgrass	1.4	1	Ensembl Plants
Linum usitatissimum	Flax	1	1	DOE-JGI; Phytozome 10
Malus domestica	Apple	1	1	DOE-JGI; Phytozome 10
Manihot esculenta	Cassava	6	6.1	DOE-JGI; Phytozome 10
Medicago truncatula	Barrel medic	Mt4.0	Mt4.0v1	DOE-JGI; Phytozome 10
Mimulus guttatus	Monkey flower	2	2	DOE-JGI; Phytozome 10
Musa acuminata	Banana	MA1	MA1	Ensembl Plants
Oryza sativa	Asian rice	IRGSP-1.0	7	DOE-JGI; Phytozome 10
Panicum virgatum	Switchgrass	1	1.1	DOE-JGI; Phytozome 10
Phaseolus vulgaris	Common bean	1	1	DOE-JGI; Phytozome 10
Populus trichocarpa	Western poplar	3	3	DOE-JGI; Phytozome 10
Prunus persica	Peach	2	2.1	DOE-JGI; Phytozome 10
Ricinus communis	Castor bean	0.1	0.1	DOE-JGI; Phytozome 10
Setaria italica	Foxtail millet	2	2.1	DOE-JGI; Phytozome 10
Solanum lycopersicum	Tomato	SL2.50	iTAG2.3	DOE-JGI; Phytozome 10
Solanum tuberosum	Potato	3_2.1.10	3.4	DOE-JGI; Phytozome 10
Sorghum bicolor	Cereal grass	2	2.1	DOE-JGI; Phytozome 10
Theobroma cacao	Cacao	1	1	DOE-JGI; Phytozome 10
Triticum urartu	Red wild einkorn	ASM34745v1	1	Ensembl Plants
Vitis vinifera	Grape	Genoscope.12X	Genoscope.12X	DOE-JGI; Phytozome 10

Zea mays	Maize	6a	6a	DOE-JGI; Phytozome 10

Table S2: List of all species and genome assembly versions, annotation versions, and data sources for sequences used in the likelihood ratio test.

Name	Region of Origin	Date Released	Use	Row Type	Growth Habit
Barke	Germany	1996	Malt	Two Row	Spring
Borwina	Germany	1983	—	Six Row	Winter
Bonus	Sweden	1959	—	Two Row	Spring
Bowman	North Dakota	1985	Feed	Two Row	Spring
Foma	Sweden	1964	—	Two Row	Spring
Gull	Sweden	1897	—	Two Row	Spring
Harrington	Canada	1981	Malting	Two Row	Spring
Haruna Nijo	Japan	1981	Malting	Two Row	Spring
Igri	Germany	1978	—	Two Row	Winter
Kindred	North Dakota	1958	Malting	Six Row	Spring
Morex	Minnesota	1979	Malting	Six Row	Spring
Steptoe	Washington State	1973	Feed	Six Row	Spring
Vogelsanger Gold	Germany	1965	—	Two Row	Spring
FT11	Wild	—	—	Two Row	—
OUH602	Eastern Wild	—	—	Two Row	—

Table S3: Location and date of cultivar release for the barley accessions used in this study.

Name	Region of Origin	Date Released
Archer	Iowa	1991
Minsoy	Minnesota	1954
Noir1	France	1963
Williams 82	Illinois	1988
M92-220	Minnesota	2008?
IA 3023	Iowa	—
Williams	Illinois	1970

Table S4: Location and date of cultivar release for the soybean accessions used in this study

Sample	Differences From Morex	Noncoding	Synonymous	Non-synonymous	Non-sense
Barke	162954	130832	17800	14322	86
Bonus	135540	108762	15184	11594	65
Borwina	139222	111613	15699	11910	62
Bowman	130335	105575	13645	11115	77
Foma	156846	126608	16746	13492	84
Gull	128671	103143	14555	10973	55
Harrington	153203	124264	15995	12944	89
Haruna Nijo	155245	125143	16655	13447	83
Igri	161224	130100	17275	13849	98
Kindred	86932	69996	9458	7478	34
Morex	5239	4602	253	384	4
Steptoe	148785	120245	15839	12701	91
Vogelsanger Gold	146303	117557	16168	12578	70
FT11	243049	195009	26886	21154	147
OUH602	197405	158377	21701	17327	115

Table S5: Counts of SNPs in various classes in thirteen barley accessions and two wild barley accessions. Numbers reported are comparisons against the reference genome, which makes it possible to include noncoding variants, where ancestral state cannot be estimated unambiguously

Sample	Differences From Wm82	Noncoding	Synonymous	Non-synonymous	Non-sense
Archer	71400	38056	12397	16948	457
IA3023	60388	33905	11335	14703	426
M92-220	85861	45183	14644	19843	591
Minsoy	151622	77466	23614	31430	927
Noir1	128689	65508	20292	27309	828
Williams	3340	7308	3688	3854	87
Williams 82	9497	6,267	3,248	3,359	70
G. soja	163365	83940	24120	32114	930

Table S6: Counts of SNPs in various classes in seven soybean accessions and one wild soybean accessions. Numbers reported are comparisons against the reference genome, which makes it possible to include noncoding variants, where ancestral state cannot be estimated unambiguously.

Sample	SIFT	PPH	LRT	Intersect
Barke	3609 (0.199)	3507 (0.194)	3339 (0.185)	1070 (0.059)
Bonus	3066 (0.189)	2911 (0.18)	2941 (0.182)	882 (0.054)
Borwina	3083 (0.19)	2932 (0.181)	2977 (0.184)	894 (0.055)
Bowman	3335 (0.186)	3219 (0.18)	3248 (0.182)	998 (0.056)
Foma	3879 (0.2)	3666 (0.189)	3598 (0.186)	1151 (0.059)
Gull	3076 (0.187)	2946 (0.179)	2998 (0.183)	885 (0.054)
Harrington	3599 (0.196)	3536 (0.192)	3335 (0.182)	1063 (0.058)
Haruna Nijo	3576 (0.194)	3507 (0.191)	3354 (0.182)	1078 (0.059)
Igri	3694 (0.199)	3585 (0.193)	3350 (0.181)	1077 (0.058)
Kindred	3131 (0.175)	2995 (0.168)	3204 (0.18)	931 (0.052)
Morex	2492 (0.146)	2298 (0.135)	3034 (0.178)	700 (0.041)
Steptoe	3479 (0.195)	3393 (0.19)	3266 (0.183)	1061 (0.06)

Vogelsanger Gold	3172 (0.193)	3100 (0.188)	2995 (0.182)	910 (0.055)
FT11	3948 (0.225)	3937 (0.224)	3290 (0.187)	1220 (0.069)
OUH602	3864 (0.209)	3886 (0.21)	3391 (0.183)	1174 (0.064)
Joint	13626 (0.228)	13534 (0.226)	11574 (0.193)	4275 (0.071)

Table S7: Per-approach and per-sample counts of deleterious variants for barley. Numbers reported are comparisons against ancestral state. The proportion of nonsynonymous variants that is inferred to be deleterious by each prediction approach in each accession is shown in parentheses

Sample	SIFT	PPH	LRT	Intersect
Archer	1987 (0.062)	3847 (0.12)	3166 (0.099)	773 (0.024)
IA3023	1994 (0.062)	3837 (0.119)	3178 (0.099)	792 (0.025)
M92-220	2142 (0.066)	4269 (0.132)	3397 (0.105)	860 (0.027)
Minsoy	2686 (0.081)	5257 (0.158)	3977 (0.119)	1135 (0.034)
Noir 1	2417 (0.073)	4951 (0.15)	3865 (0.117)	1035 (0.031)
Williams	1408 (0.048)	2673 (0.09)	2425 (0.082)	485 (0.016)
Williams 82	1394 (0.047)	2631 (0.089)	2393 (0.081)	478 (0.016)
G. soja	1751 (0.074)	3583 (0.15)	2675 (0.112)	716 (0.03)
Joint	7694 (0.076)	14933 (0.147)	11223 (0.11)	3041 (0.03)

Table S8: Per-approach and per-sample of counts of deleterious variants in soybean. Numbers reported are comparisons against ancestral state. The proportion of nonsynonymous variants that is inferred to be deleterious by each prediction approach in each accession is shown in parentheses.

Gene Name	Org.	Gene ID	Tx. ID	AA1	A2	CDS Pos.	Chr.	Pos.	SNP ID	Samp.	Caus.	SIFT	PPH	LRT	Ref.
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	P	S	52	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	D	N	71	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	D	N	73	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	S	N	78	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	R	W	83	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	P	L	113	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	G	D	116	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	P	A	135	morex_contig_274284	918	Barley_379226	Y	Y	T	T	D	Comadran et al., 2012
CENTRO-RADIALIS	Barley	MLOC_44160	MLOC_44160.1	R	W	139	morex_contig_274284	-	-	N	N	D	D	D	Comadran et al., 2012
Vrs1	Barley	MLOC_4993	MLOC_4993.2	D	G	8	morex_contig_135757	-	-	N	N	T	NA	NA	Taketa et al., 2007
Vrs1	Barley	MLOC_4993	MLOC_4993.2	D	E	26	morex_contig_135757	-	-	N	N	NA	NA	NA	Taketa et al., 2007

Vrs1	Barley	MLOC_4993	MLOC_4993.2	F	A	75	morex_contig_135757	-	-	N	Y	T	NA	NA	Taketa et al., 2007
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	Q	H	14	morex_contig_94710	4424	Barley_237670	Y	N	T	T	T	Jones et al., 2008
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	K	R	285	morex_contig_94710	2690	Barley_237653	Y	N	T	T	T	Jones et al., 2008
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	P	L	303	morex_contig_94710	2636	Barley_237652	Y	N	T	T	T	Jones et al., 2008
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	N	K	304	morex_contig_94710	2632	Barley_237651	Y	N	T	T	T	Jones et al., 2008
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	S	P	344	morex_contig_94710	2514	Barley_237650	Y	N	T	T	T	Jones et al., 2008
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	S	P	369	morex_contig_94710	2440	Barley_237647	Y	Y	T	T	T	Jones et al., 2008
Ppd-H1	Barley	MLOC_81154	MLOC_81154.1	P	L	399	morex_contig_94710	2348	Barley_652697	Y	N	D	T	T	Jones et al., 2008
BAM4	Barley	MLOC_4517	MLOC_4517.1	V	E	85	morex_contig_135213	-	-	N	Y	D	D	D	Eglinton et al., 1998
BAM4	Barley	MLOC_4517	MLOC_4517.1	A	V	233	morex_contig_135213	-	-	N	Y	T	T	NA	Eglinton et al., 1998
BAM4	Barley	MLOC_4517	MLOC_4517.1	R	L	347	morex_contig_135213	-	-	N	Y	D	D	D	Eglinton et al., 1998
BAM4	Barley	MLOC_4517	MLOC_4517.1	R	S	347	morex_contig_135213	-	-	N	Y	D	D	D	Eglinton et al., 1998
INT-C	Barley	MLOC_70116	MLOC_70116.1	A	P	42	morex_contig_5747	4964	Barley_1587	Y	N	T	T	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	S	P	79	morex_contig_5747	5075	Barley_1589	Y	N	T	T	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	A	V	95	morex_contig_5747	5124	Barley_1590	Y	N	T	T	T	Ramsay et al., 2011

INT-C	Barley	MLOC_70116	MLOC_70116.1	R	V	131	morex_contig_5747	-	-	N	Y*	T	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	S	L	137	morex_contig_5747	-	-	N	Y*	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	A	V	140	morex_contig_5747	-	-	N	Y*	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	F	I	144	morex_contig_5747	-	-	N	Y*	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	L	F	146	morex_contig_5747	-	-	N	Y	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	Q	L	147	morex_contig_5747	-	-	N	Y*	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	W	R	162	morex_contig_5747	-	-	N	Y*	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	L	R	164	morex_contig_5747	-	-	N	Y*	NA	NA	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	S	N	233	morex_contig_5747	5538	Barley_1592	Y	N	T	T	D	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	A	T	249	morex_contig_5747	-	-	N	Y	T	NA	T	Ramsay et al., 2011
INT-C	Barley	MLOC_70116	MLOC_70116.1	R	G	254	morex_contig_5747	-	-	N	Y	T	NA	T	Ramsay et al., 2011
Rpg1	Barley	MLOC_15351	MLOC_15351.1	V	E	91	morex_contig_1570478	4129	Barley_477718	Y	N	T	T	T	Bruggeman et al., 2002
Rpg1	Barley	MLOC_15351	MLOC_15351.1	G	E	159	morex_contig_1570478	3678	Barley_752463	Y	N	T	T	T	Bruggeman et al., 2002
Rpg1	Barley	MLOC_15351	MLOC_15351.1	Q	*	192	morex_contig_1570478	3580	Barley_477717	Y	Y	D	D	D	Bruggeman et al., 2002
Rpg1	Barley	MLOC_15351	MLOC_15351.1	N	S	198	morex_contig_1570478	3561	Barley_752461	Y	N	T	T	T	Bruggeman et al., 2002

Rpg1	Barley	MLOC_15351	MLOC_15351.1	K	Q	436	morex_contig_1570478	3659	Barley_477716	Y	Y	T	T	T	Bruggeman et al., 2002
Rpg1	Barley	MLOC_15351	MLOC_15351.1	I	V	439	morex_contig_1570478	2560	Barley_752455	Y	N	T	T	T	Bruggeman et al., 2002
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	Q	149	morex_contig_137037	4539	Barley_280801	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	Y	156	morex_contig_137037	4561	Barley_280802	Y	N	T	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	T	174	morex_contig_137037	4615	Barley_671267	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	E	177	morex_contig_137037	4625	Barley_280809	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	I	180	morex_contig_137037	4633	Barley_280810	Y	N	T	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	P	H	183	morex_contig_137037	4642	Barley_280811	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	V	I	188	morex_contig_137037	4656	Barley_671268	Y	N	D	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	A	V	191	morex_contig_137037	4666	Barley_280813	Y	N	D	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	G	D	193	morex_contig_137037	4672	Barley_671270	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	H	197	morex_contig_137037	4684	Barley_280816	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	R	198	morex_contig_137037	4687	Barley_671272	Y	N	D	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	N	198	morex_contig_137037	4688	Barley_671273	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	E	206	morex_contig_137037	4710	Barley_671274	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	M	L	208	morex_contig_137037	4716	Barley_280817	Y	N	T	T	T	Wang et al., 2013

RGA1	Barley	MLOC_6270	MLOC_6270.1	M	K	208	morex_contig_137037	4717	Barley_671276	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	S	209	morex_contig_137037	4720	Barley_280819	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	T	211	morex_contig_137037	4726	Barley_671277	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	T	A	213	morex_contig_137037	4731	Barley_280822	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	N	215	morex_contig_137037	4737	Barley_280823	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	A	218	morex_contig_137037	4747	Barley_280824	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	N	S	222	morex_contig_137037	4759	Barley_280825	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	A	S	290	morex_contig_137037	6389	Barley_280831	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	T	I	292	morex_contig_137037	6396	Barley_280833	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	T	295	morex_contig_137037	6405	Barley_280834	Y	N	T	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	R	308	morex_contig_137037	6445	Barley_280838	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	Q	314	morex_contig_137037	6461	Barley_280842	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	M	315	morex_contig_137037	6464	Barley_280843	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	T	A	316	morex_contig_137037	6467	Barley_280844	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	A	323	morex_contig_137037	6488	Barley_280846	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	M	332	morex_contig_137037	6515	Barley_280849	Y	N	D	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	R	G	336	morex_contig_137037	6527	Barley_280851	Y	N	D	T	D	Wang et al., 2013

RGA1	Barley	MLOC_6270	MLOC_6270.1	R	H	336	morex_contig_137037	6528	Barley_280852	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	N	339	morex_contig_137037	6537	Barley_280856	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	N	340	morex_contig_137037	-	-	N	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	I	V	342	morex_contig_137037	6545	Barley_280858	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	I	M	342	morex_contig_137037	6547	Barley_280859	Y	N	D	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	M	I	357	morex_contig_137037	6592	Barley_280863	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	T	A	385	morex_contig_137037	6674	Barley_671280	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	G	A	445	morex_contig_137037	6855	Barley_280880	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	G	S	446	morex_contig_137037	6857	Barley_280882	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	R	451	morex_contig_137037	6873	Barley_280885	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	P	453	morex_contig_137037	6879	Barley_280886	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	P	A	454	morex_contig_137037	6881	Barley_280887	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	M	D	474	morex_contig_137037	-	-	N	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	N	474	morex_contig_137037	6941	Barley_280891	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	E	474	morex_contig_137037	6943	Barley_280892	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	E	478	morex_contig_137037	6953	Barley_280894	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	T	A	492	morex_contig_137037	6995	Barley_280897	Y	N	T	T	T	Wang et al., 2013

RGA1	Barley	MLOC_6270	MLOC_6270.1	I	V	494	morex_contig_137037	7001	Barley_280898	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	K	505	morex_contig_137037	7034	Barley_280900	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	E	519	morex_contig_137037	7078	Barley_280903	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	M	531	morex_contig_137037	7112	Barley_280907	Y	N	D	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	V	A	551	morex_contig_137037	7173	Barley_280912	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	M	T	568	morex_contig_137037	7224	Barley_280919	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	H	Q	575	morex_contig_137037	7246	Barley_280920	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	F	586	morex_contig_137037	7277	Barley_280925	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	V	589	morex_contig_137037	7286	Barley_280926	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	H	595	morex_contig_137037	7306	Barley_671282	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	P	603	morex_contig_137037	7328	Barley_671284	Y	N	D	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	S	*	603	morex_contig_137037	7329	Barley_671285	Y	N	D	D	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	G	R	627	morex_contig_137037	7400	Barley_280933	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	R	C	628	morex_contig_137037	7403	Barley_280935	Y	N	D	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	V	G	630	morex_contig_137037	7410	Barley_280936	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	V	649	morex_contig_137037	7467	Barley_671288	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	M	T	655	morex_contig_137037	7485	Barley_671290	Y	N	T	T	D	Wang et al., 2013

RGA1	Barley	MLOC_6270	MLOC_6270.1	M	I	655	morex_contig_137037	7486	Barley_671291	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	R	G	660	morex_contig_137037	7499	Barley_671292	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	G	661	morex_contig_137037	7503	Barley_671293	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	D	661	morex_contig_137037	7504	Barley_671294	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	L	M	663	morex_contig_137037	7508	Barley_671295	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	N	664	morex_contig_137037	7511	Barley_671296	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	N	669	morex_contig_137037	7526	Barley_671299	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	M	I	674	morex_contig_137037	7543	Barley_671301	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	G	E	678	morex_contig_137037	7554	Barley_671302	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	R	G	686	morex_contig_137037	7577	Barley_280938	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	R	Q	686	morex_contig_137037	7578	Barley_280939	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	T	K	694	morex_contig_137037	7602	Barley_280940	Y	N	D	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	R	708	morex_contig_137037	7644	Barley_280942	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	L	711	morex_contig_137037	7653	Barley_671303	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	R	713	morex_contig_137037	7659	Barley_671304	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	Q	H	713	morex_contig_137037	7660	Barley_671305	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	N	S	728	morex_contig_137037	7704	Barley_671310	Y	N	T	T	T	Wang et al., 2013

RGA1	Barley	MLOC_6270	MLOC_6270.1	D	N	736	morex_contig_137037	7727	Barley_671313	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	D	G	736	morex_contig_137037	7728	Barley_671314	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	G	740	morex_contig_137037	7740	Barley_671315	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	R	754	morex_contig_137037	7782	Barley_671318	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	G	756	morex_contig_137037	7788	Barley_671319	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	E	D	756	morex_contig_137037	7789	Barley_671320	Y	N	T	T	D	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	N	D	804	morex_contig_137037	7931	Barley_280954	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	H	R	812	morex_contig_137037	7956	Barley_280956	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	R	G	843	morex_contig_137037	8048	Barley_280957	Y	N	T	D	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	V	G	866	morex_contig_137037	8118	Barley_280958	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	N	873	morex_contig_137037	8140	Barley_280959	Y	N	T	T	T	Wang et al., 2013
RGA1	Barley	MLOC_6270	MLOC_6270.1	K	R	878	morex_contig_137037	8154	Barley_280960	Y	N	T	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	Y	85	morex_contig_48606	5347	Barley_126528	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	A	86	morex_contig_48606	5349	Barley_126529	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	P	S	88	morex_contig_48606	5355	Barley_126531	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	G	V	92	morex_contig_48606	5368	Barley_126532	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	E	D	95	morex_contig_48606	5378	Barley_126533	Y	N	T	T	T	Wang et al., 2013

RPG5	Barley	MLOC_64296	MLOC_64296.1	C	R	101	morex_contig_48606	5394	Barley_126535	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	I	104	morex_contig_48606	5404	Barley_126536	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	P	A	106	morex_contig_48606	5409	Barley_126537	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	V	L	108	morex_contig_48606	5415	Barley_126538	Y	N	D	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	D	Y	112	morex_contig_48606	5427	Barley_126541	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	D	G	112	morex_contig_48606	5428	Barley_126542	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	R	G	113	morex_contig_48606	5430	Barley_126544	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	Q	E	118	morex_contig_48606	5445	Barley_126545	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	D	E	119	morex_contig_48606	5450	Barley_126546	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	P	A	129	morex_contig_48606	5478	Barley_608197	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	P	R	129	morex_contig_48606	5479	Barley_126547	Y	N	T	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	I	V	130	morex_contig_48606	5481	Barley_608198	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	E	D	131	morex_contig_48606	5486	Barley_608199	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	N	H	132	morex_contig_48606	5487	Barley_608200	Y	N	D	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	V	D	135	morex_contig_48606	5497	Barley_608201	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	L	M	137	morex_contig_48606	5502	Barley_608202	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	*	142	morex_contig_48606	5518	Barley_608204	Y	Y	D	D	D	Wang et al., 2013

RPG5	Barley	MLOC_64296	MLOC_64296.1	S	P	142	morex_contig_48606	5517	Barley_608203	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	H	D	145	morex_contig_48606	5526	Barley_608206	Y	N	D	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	N	S	148	morex_contig_48606	5536	Barley_608208	Y	N	T	T	D	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	M	L	155	morex_contig_48606	5556	Barley_608209	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	A	S	157	morex_contig_48606	5562	Barley_608210	Y	N	T	T	D	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	F	S	161	morex_contig_48606	5575	Barley_608211	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	T	169	morex_contig_48606	5599	Barley_608212	Y	N	D	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	A	T	236	morex_contig_48606	5799	Barley_126549	Y	N	T	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	R	K	265	morex_contig_48606	5887	Barley_126551	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	T	S	267	morex_contig_48606	5893	Barley_126552	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	F	L	273	morex_contig_48606	5912	Barley_126553	Y	N	D	D	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	N	672	morex_contig_48606	7265	Barley_126555	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	L	V	767	morex_contig_48606	8207	Barley_126556	Y	N	T	T	T	Wang et al., 2013
RPG5	Barley	MLOC_64296	MLOC_64296.1	S	A	771	morex_contig_48606	8219	Barley_126557	Y	N	T	T	D	Wang et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	Q	H	125	morex_contig_106547	3765	Barley_656577	Y	N	D	D	D	Nishida et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	F	S	203	morex_contig_106547	3998	Barley_246823	Y	Y	D	D	D	Nishida et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	G	S	472	morex_contig_106547	4804	Barley_656578	Y	N	T	T	D	Nishida et al., 2013

PhyC	Barley	MLOC_824	MLOC_824.1	T	I	679	morex_contig_106547	5536	Barley_656580	Y	N	T	D	D	Nishida et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	R	C	819	morex_contig_106547	6303	Barley_656582	Y	N	D	T	D	Nishida et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	V	A	825	morex_contig_106547	6322	Barley_656583	Y	N	D	T	D	Nishida et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	Q	R	875	morex_contig_106547	6472	Barley_656584	Y	N	D	T	T	Nishida et al., 2013
PhyC	Barley	MLOC_824	MLOC_824.1	A	T	946	morex_contig_106547	7111	Barley_246827	Y	N	T	T	T	Nishida et al., 2013
ELF3	Barley	MLOC_75281	MLOC_75281.1	R	H	43	morex_contig_67536	-	-	N	Y	T	NA	T	Zakhrabekova et al., 2012
ELF3	Barley	MLOC_75281	MLOC_75281.1	N	I	44	morex_contig_67536	-	-	N	Y	T	NA	T	Zakhrabekova et al., 2012
ELF3	Barley	MLOC_75281	MLOC_75281.1	K	E	276	morex_contig_67536	1653	Barley_207366	Y	N	T	T	T	Zakhrabekova et al., 2012
ELF3	Barley	MLOC_75281	MLOC_75281.1	G	W	279	morex_contig_67536	1644	Barley_207365	Y	N	D	D	D	Zakhrabekova et al., 2012
ELF3	Barley	MLOC_75281	MLOC_75281.1	P	L	308	morex_contig_67536	1556	Barley_641053	Y	N	T	T	D	Zakhrabekova et al., 2012
Dt1	Soybean	GLYMA19G37890	GLYMA19G37890.1	R	W	166	Gm19	-	-	N	Y	D	D	D	Liu et al., 2010
ChII1a	Soybean	GLYMA13G30560	GLYMA13G30560.1	R	Q	273	Gm13	-	-	N	Y	D	D	D	Campbell et al., 2015
ChII1a	Soybean	GLYMA13G30560	GLYMA13G30560.1	Q	R	275	Gm13	-	-	N	Y	D	D	D	Campbell et al., 2015

FAD2-1B	Soybean	GLYMA 20G24530	GLYMA20G24530.1	P	R	137	Gm20	-	-	N	Y	D	D	D	Pham et al., 2010
FAD2-1B	Soybean	GLYMA 20G24530	GLYMA20G24530.1	I	T	143	Gm20	-	-	N	N	D	T	D	Pham et al., 2010
GMPHYA3	Soybean	GLYMA 19G41210	GLYMA19G41210.1	G	R	1050	Gm19	-	-	N	Y	T	D	NA	Watanabe et al., 2009
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	H	A	15	Gm16	-	-	N	N	D	D	D	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	D	A	125	Gm16	-	-	N	N	D	D	D	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	H	A	359	Gm16	-	-	N	N	D	D	T	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	H	A	368	Gm16	-	-	N	N	D	D	D	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	E	A	376	Gm16	-	-	N	N	D	D	T	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	E	A	392	Gm16	-	-	N	Y	D	D	D	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	H	D	392	Gm16	-	-	N	Y	D	NA	D	Noguchi et al., 2007
GmIF7GT	Soybean	GLYMA 16G29400	GLYMA16G29400.1	H	A	456	Gm16	-	-	N	N	D	NA	D	Noguchi et al., 2007

E1	Soybean	GLYMA 06G23026	GLYMA06G23026.1	S	F	17	Gm06	-	-	N	Y	T	NA	T	Xia et al., 2011
E1	Soybean	GLYMA 06G23026	GLYMA06G23026.1	R	K	15	Gm06	-	-	N	Y	T	T	T	Xia et al., 2011
E1	Soybean	GLYMA 06G23026	GLYMA06G23026.1	T	I	65	Gm06	-	-	N	Y	T	NA	T	Xia et al., 2011
W1	Soybean	GLYMA 13G04210	GLYMA13G04210.1	V	M	210	Gm13	-	-	N	Y	NA	D	D	Takahashi et al., 2010
W1	Soybean	GLYMA 13G04210	GLYMA13G04210.1	E	V	475	Gm13	-	-	N	N	NA	NA	NA	Takahashi et al., 2010

Table S9: List of cloned genes with SNPs causing phenotypic differences, and predictions for each SNP. Causative SNPs annotate as deleterious with a higher frequency than the genomic average.

SUPPLEMENTARY FIGURES

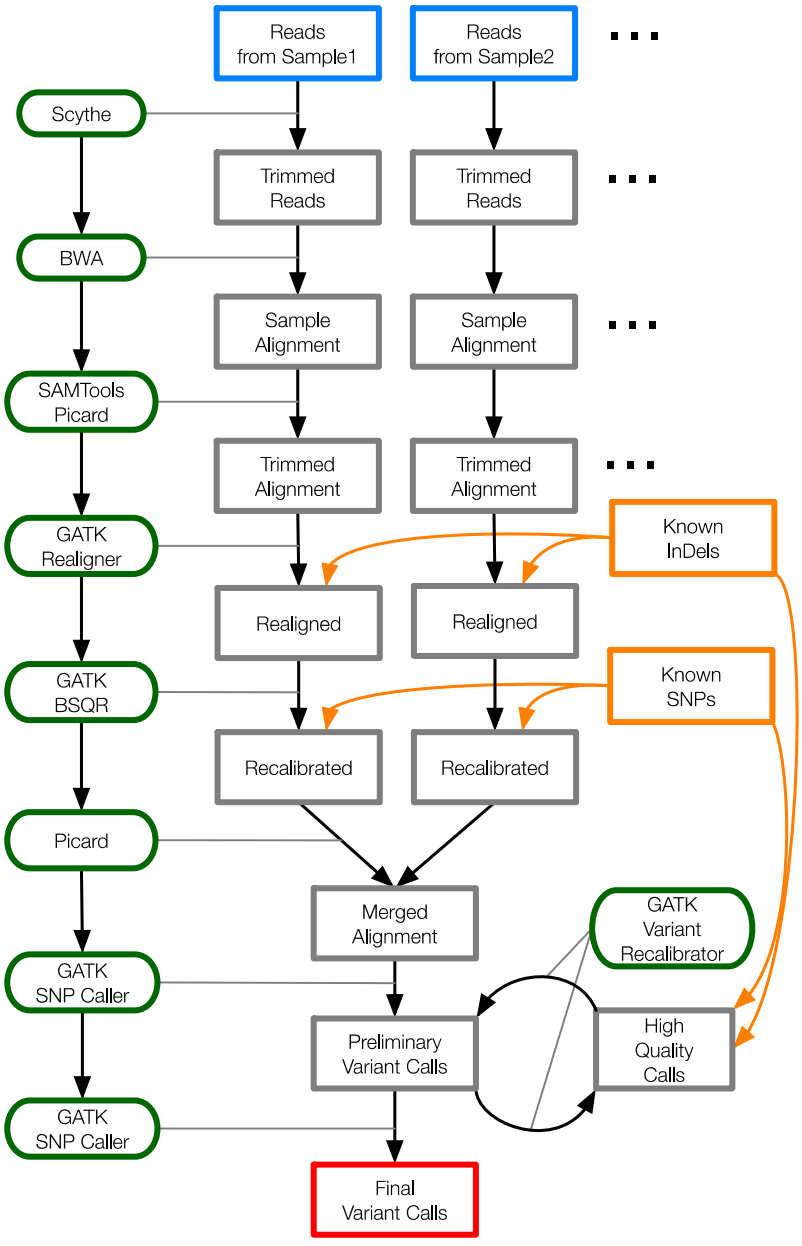

Figure S1: A schematic for the read mapping and SNP calling workflow. Boxes with bold

borders denote the start and end points of workflow. Rounded boxes with light grey borders are the tools that are used at each step in the pipeline.

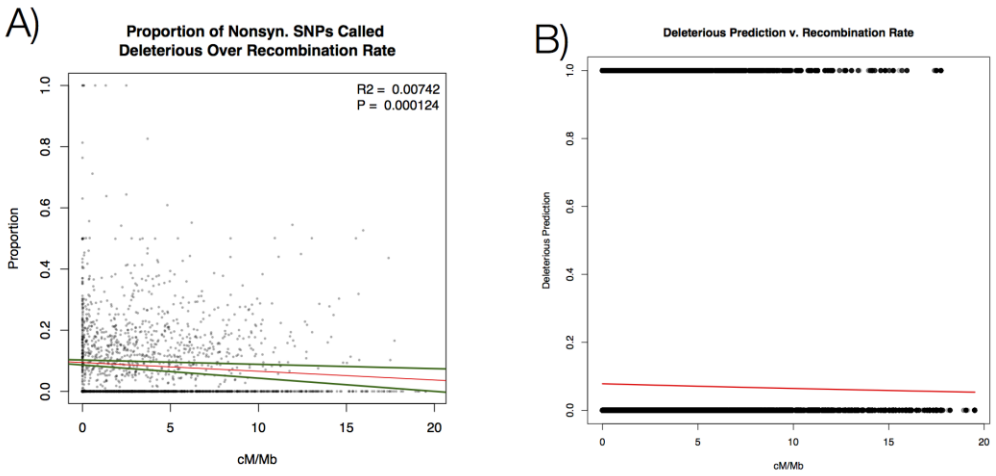

Figure S2: A) Correlation between recombination rate (cM/Mb) and proportion of nonsynonymous SNPs inferred to be deleterious genome-wide in our soybean sample. B) Logistic regression of whether or not a SNP is predicted to be deleterious against recombination rate in cM/Mb.

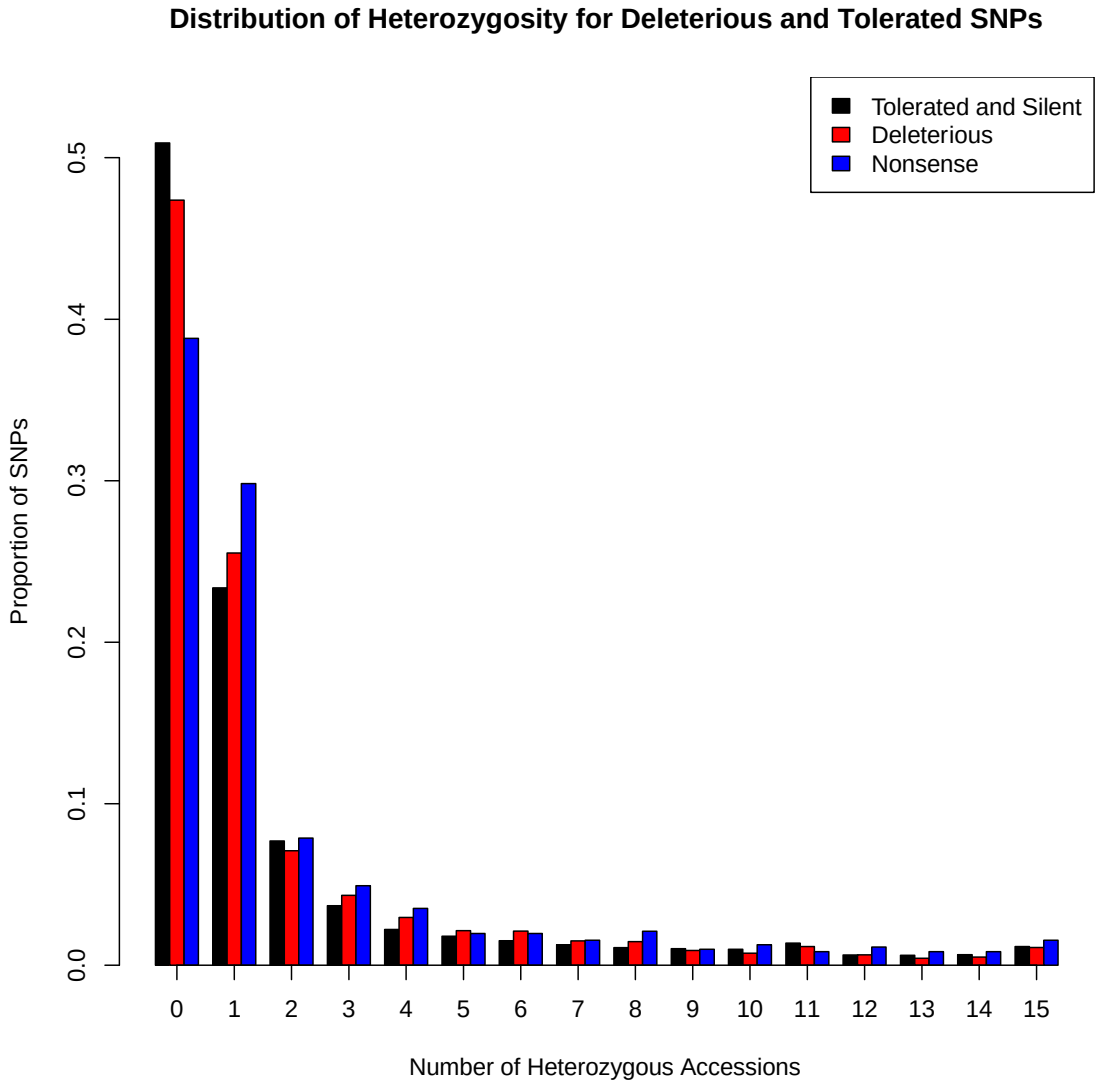

Figure S3: The distributions of per-SNP heterozygosity for tolerated nonsynonymous and silent SNPs, deleterious missense SNPs, and nonsense SNPs. Nonsense SNPs tend to be heterozygous more often than deleterious or tolerated SNPs.

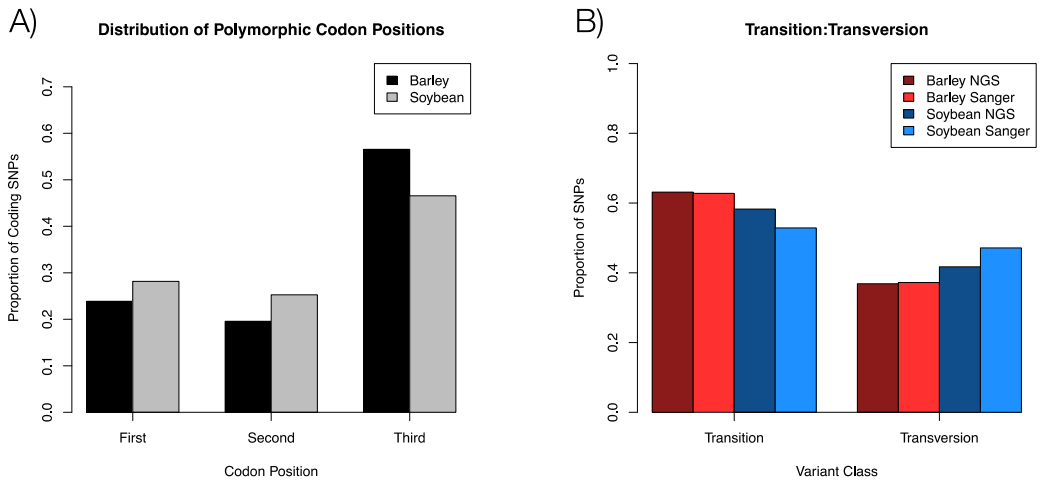

Figure S4: A) Histogram showing the relative frequencies of 1st, 2nd, and 3rd position variants in codons. The distribution shows a strong bias against 1st and 2nd positions (which tend to be nonsynonymous), consistent with the action of purifying selection. B) Proportions of SNPs identified in our datasets that are transitions and transversions. The proportions estimated from Sanger resequencing come from Morrell et al. (2006) for barley, and Hyten et al. (2006) for soybean.

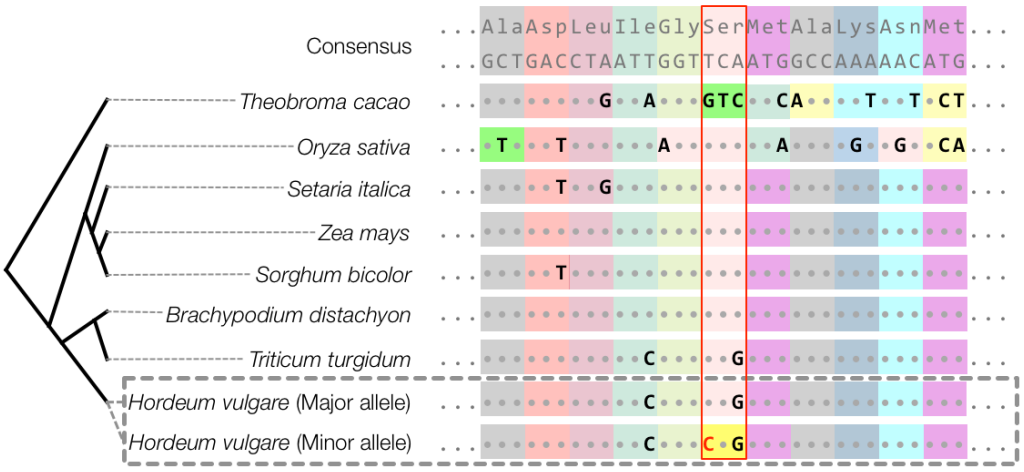

Figure S5: A sample alignment used to infer a Serine to Proline mutation as deleterious in *Ppd-H1*. The alignment is built from sequences used by SIFT, and the affected codon is highlighted in red.

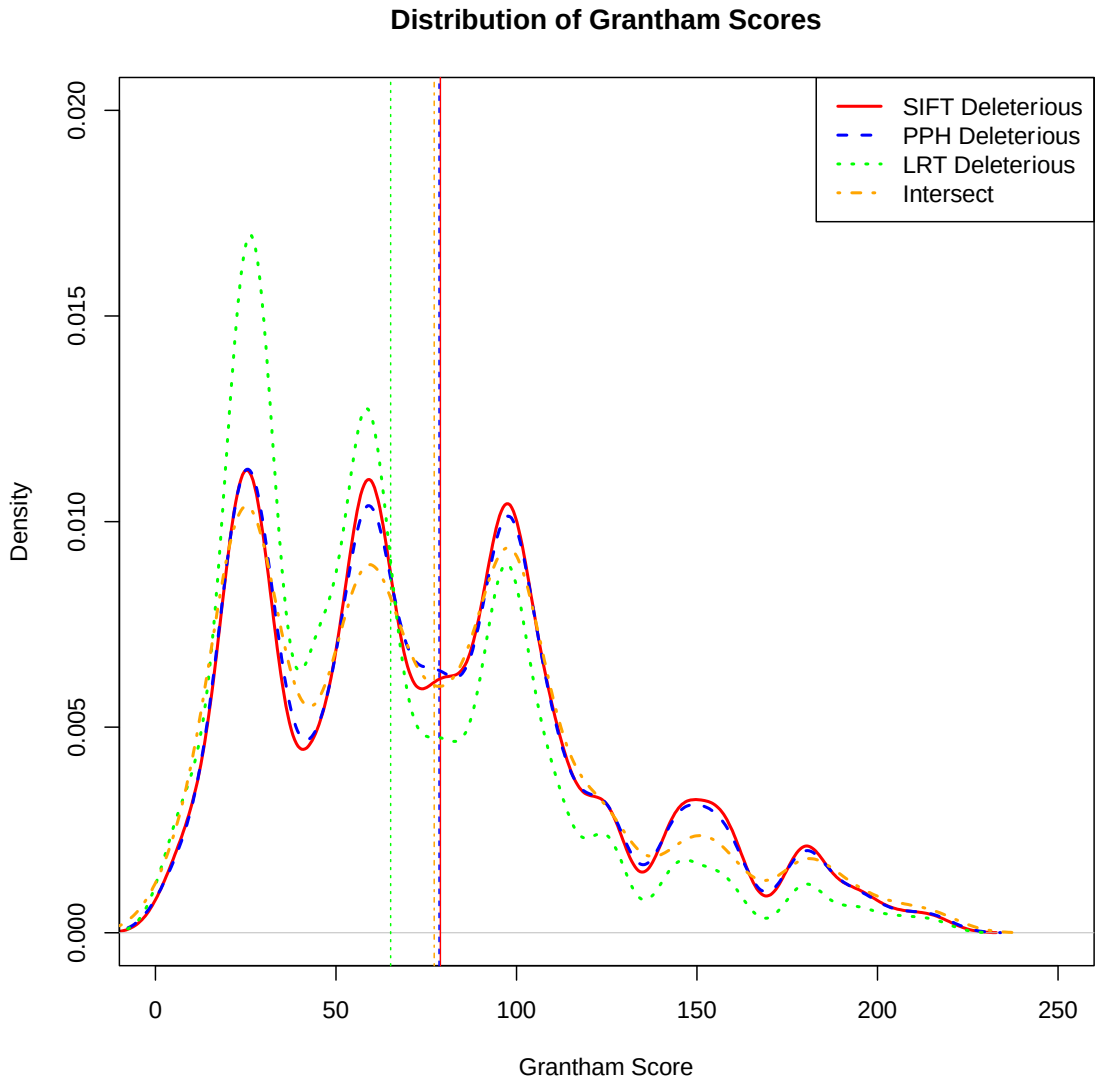

Figure S6: Distribution of Grantham score for nonsynonymous variants predicted to be deleterious by each prediction approach. Each approach and the intersection of each approach gives a very similar distribution of Grantham scores. Vertical lines show the mean of the distribution.

Appendix 3: Chapter 4 Supplementary Material

SUPPLEMENTARY FIGURES

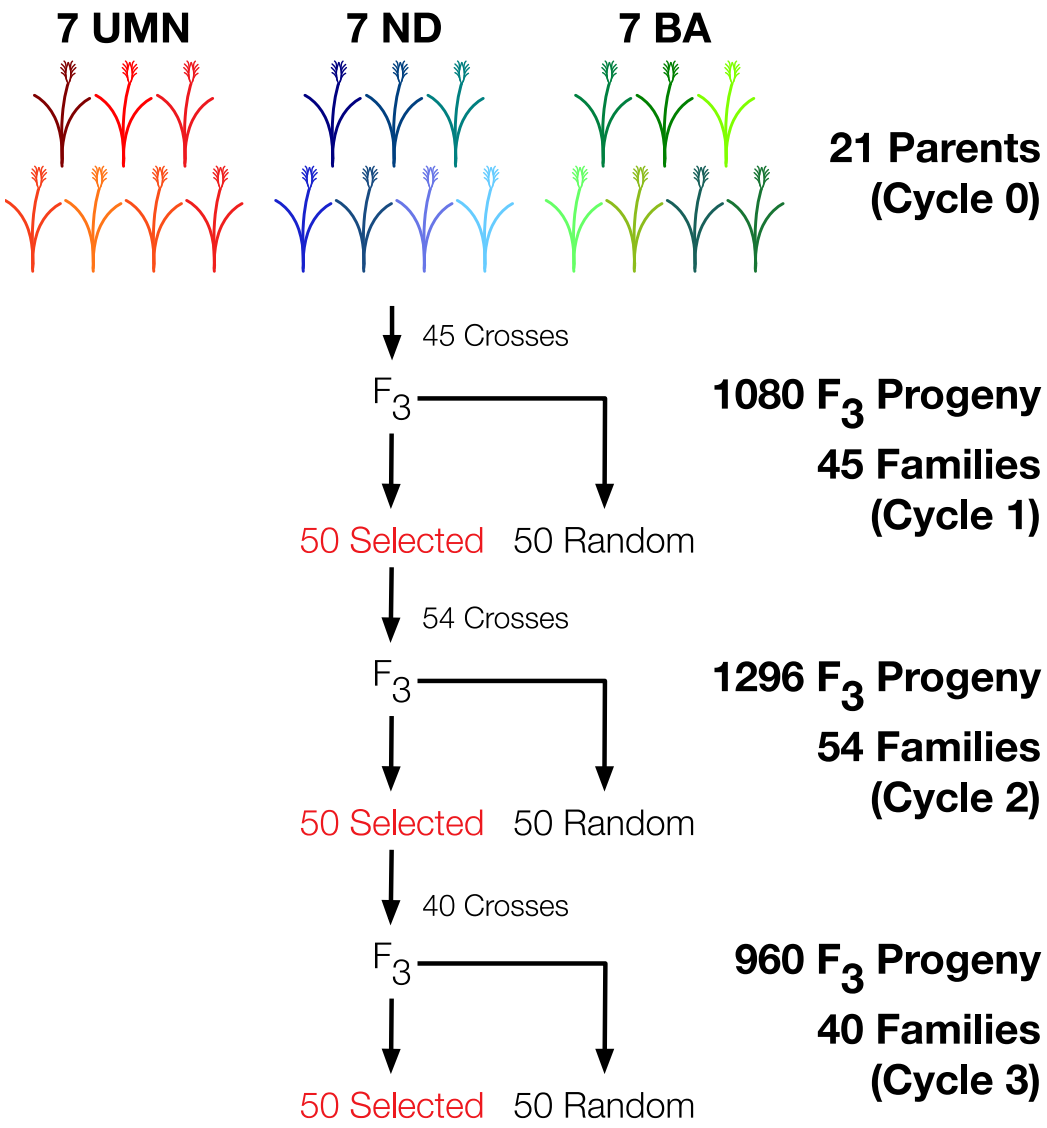

Figure S1: A schematic of the population used in this study.

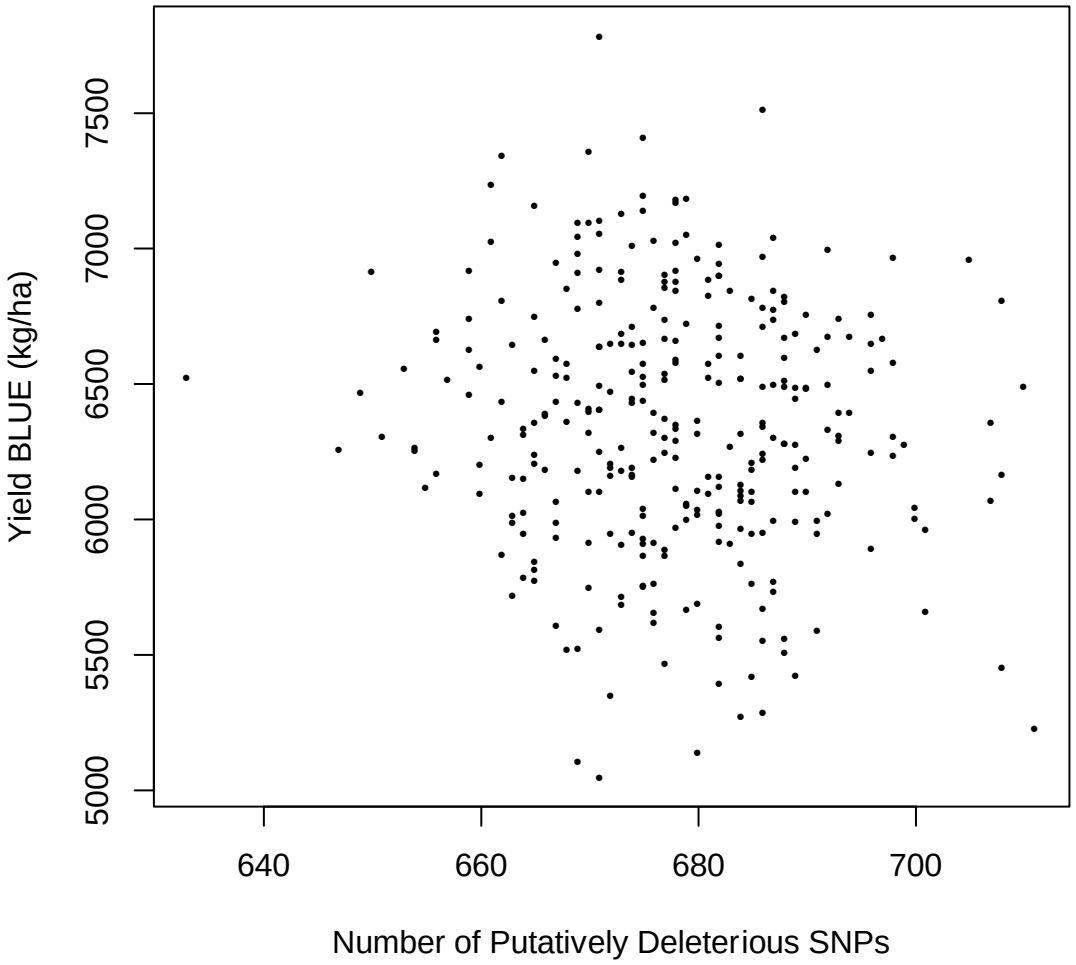

Figure S2: Plot showing the relationship between the number of putatively deleterious SNPs an individual line carries and its estimated yield. The correlation is not statistically significant ($r=-0.097$, $p > 0.01$).

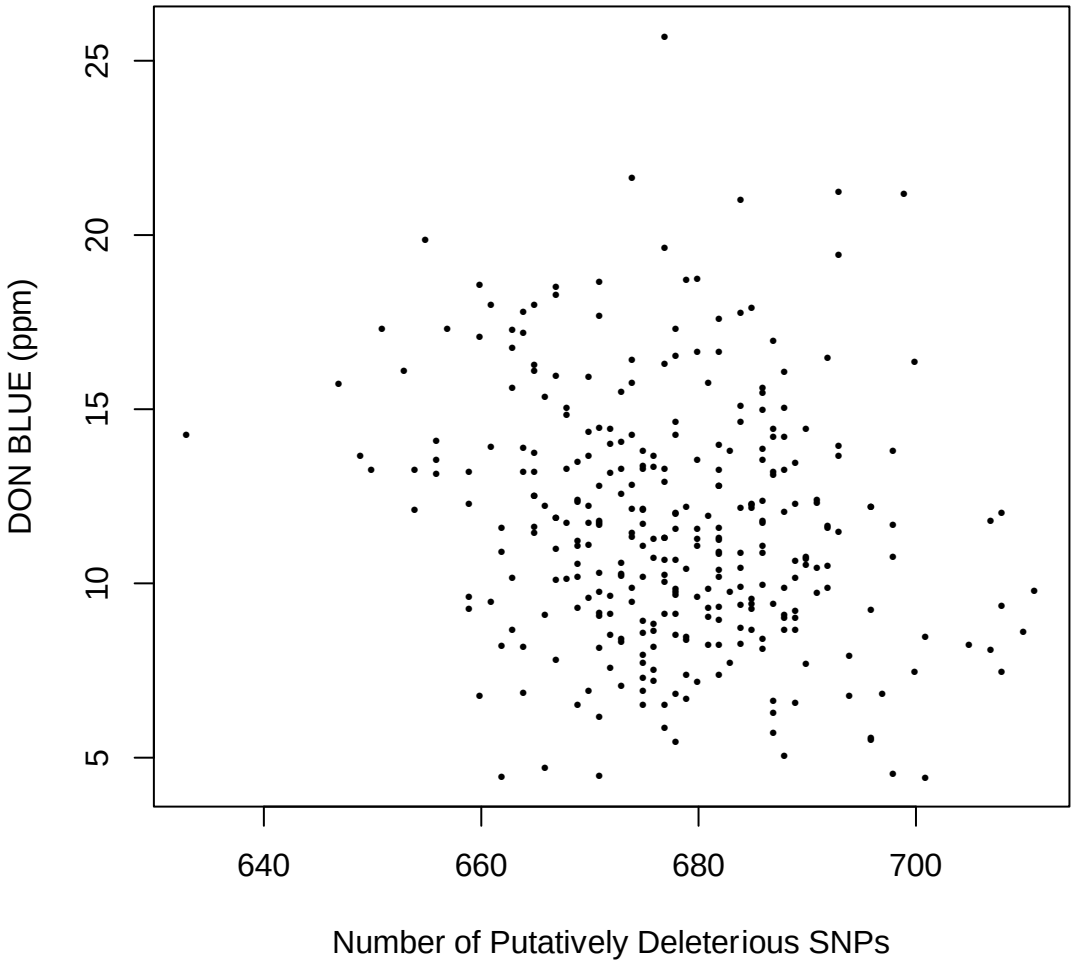

Figure S3: Relationship between the number of putatively deleterious SNPs an individual line carries and its estimated DON concentration. The correlation is negative and statistically significant ($r=-0.199$, $p < 0.01$).

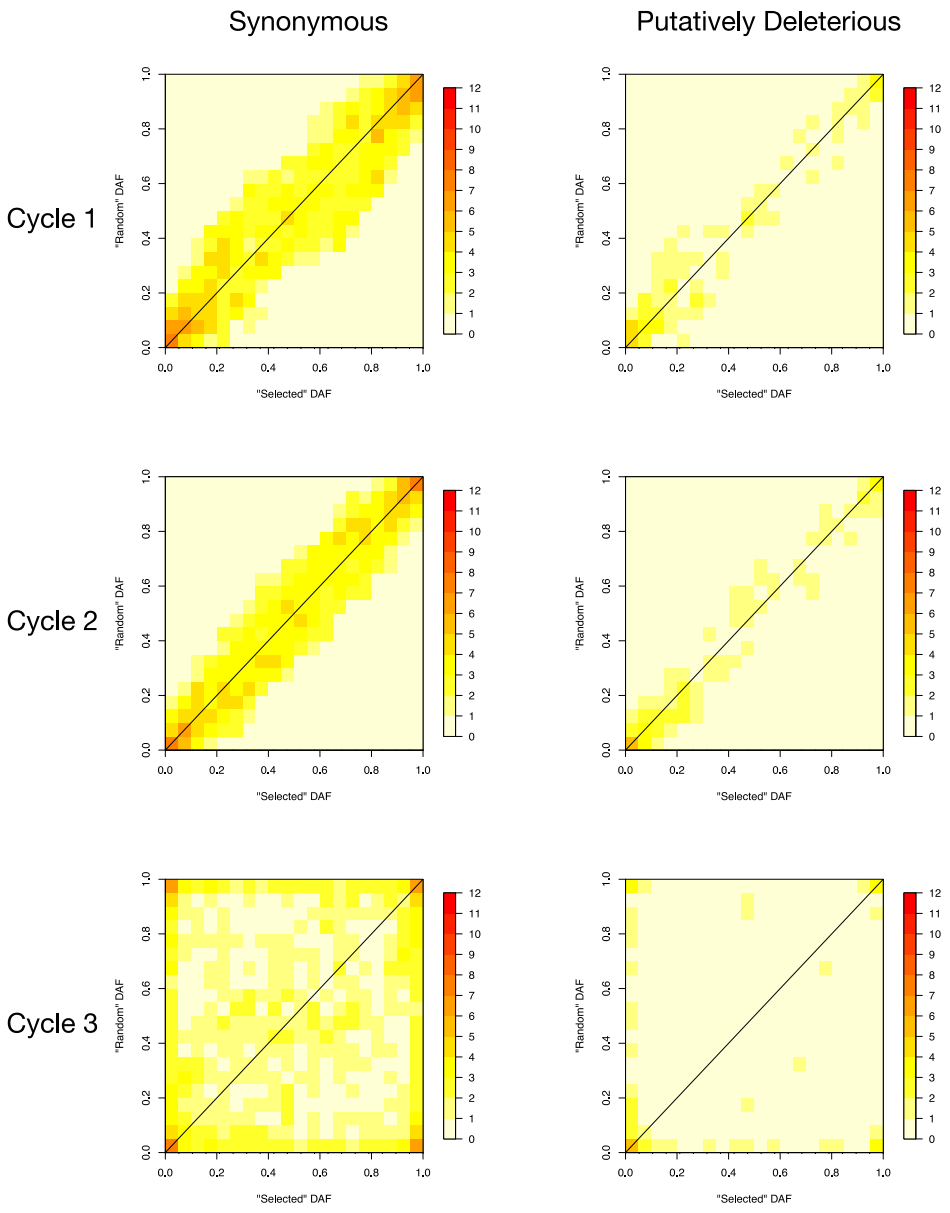

Figure S4: Joint derived site frequency spectra showing the frequencies of synonymous (putatively neutral) and putatively deleterious SNPs in the randomly chosen panels and selected panels. The color scale corresponds to the log of the number of SNPs in the corresponding frequency classes. The diagonal line shows a perfect correspondence between the frequency in the random and selected panels. Putatively deleterious SNPs do not respond differently to selection than putatively neutral SNPs.