G3 Genes | Genomes | Genetics

INVESTIGATION

Comparative Genomics Approaches Accurately Predict Deleterious Variants in Plants

Thomas J.Y. Kono,* Li Lei,* Ching-Hua Shih,† Paul J. Hoffman,* Peter L. Morrell,*,1 and Justin C. Fay†,1,2

*Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN 551085 and †Department of Genetics, Washington University, St. Louis, MO 63110

ORCID IDs: 0000-0002-1388-3194 (T.J.Y.K.); 0000-0001-5708-0118 (L.L.); 0000-0002-6652-1197 (C.-H.S.); 0000-0002-7693-8957 (P.J.H.); 0000-0001-6282-1582 (P.L.M.); 0000-0003-1893-877X (J.C.F.)

ABSTRACT Recent advances in genome resequencing have led to increased interest in prediction of the functional consequences of genetic variants. Variants at phylogenetically conserved sites are of particular interest, because they are more likely than variants at phylogenetically variable sites to have deleterious effects on fitness and contribute to phenotypic variation. Numerous comparative genomic approaches have been developed to predict deleterious variants, but the approaches are nearly always assessed based on their ability to identify known disease-causing mutations in humans. Determining the accuracy of deleterious variant predictions in nonhuman species is important to understanding evolution, domestication, and potentially to improving crop quality and yield. To examine our ability to predict deleterious variants in plants we generated a curated database of 2,910 *Arabidopsis thaliana* mutants with known phenotypes. We evaluated seven approaches and found that while all performed well, their relative ranking differed from prior benchmarks in humans. We conclude that deleterious mutations can be reliably predicted in *A. thaliana* and likely other plant species, but that the relative performance of various approaches does not necessarily translate from one species to another.

KEYWORDS
deleterious mutations
phenotypes
genome
training set

Dramatically increased numbers of reference genomes and whole genome resequencing data sets have facilitated the discovery of sequence variants and increased interest in the annotation of functional variants in many organisms. Functional annotation can yield insight into the genetic basis of phenotypic variation and is often a critical step in the identification of genes and variants underlying human disease (Ahituv *et al.* 2007; Cooper and Shendure 2011). In particular, interest in identifying putatively deleterious variants has increased, because these variants may contribute substantially to phenotypic variation (Manolio *et al.* 2009; Thornton *et al.* 2013). Because deleterious variants are more likely to disrupt phylogenetically conserved sites, the availability of comparative genomics data has made it possible to develop computational approaches to identifying deleterious variants genome-wide (Ng and Henikoff 2006). Although a number of approaches have been developed to identify deleterious variants within noncoding sequences (*e.g.*, Pollard *et al.* 2010; Kircher *et al.* 2014), most have focused on variants that alter the amino acid sequence of proteins (Ng and Henikoff 2006). This focus on amino acid substitutions in protein coding sequences is in part driven by the observation that amino acid-altering single nucleotide polymorphisms (SNPs) are more often associated with phenotypic variation than other classes of variants, but also because they are the most readily identifiable class of variants that are likely to have a biological impact (1000 Genomes Project Consortium *et al.* 2012; Fay 2013; Stenson *et al.* 2014).

While identification of disease-causing and potentially "actionable" genetic variants is fundamental to personalized medicine, identifying deleterious variants is also broadly relevant to understanding the genetic basis of phenotypic variation. In humans, annotation of deleterious variants improves prediction accuracy of complex traits (Dudley *et al.* 2012). For domesticated organisms, complementation of recessive

doi: https://doi.org/10.1534/g3.118.200563
Manuscript received July 8, 2018; accepted for publication August 10, 2018; published Early Online August 29, 2018.

Supplemental material available at Figshare: https://doi.org/10.25387/g3.6998387.
[1]Corresponding author: E-mail: justin.fay@rochester.edu or pmorrell@umn.edu
[2]Present address: Department of Biology, University of Rochester, 319 Hutchison Hall, RC Box # 270211, Rochester, NY 14627-0211

deleterious variants between haplotypes is thought to be one of the primary mechanisms underlying heterosis (Charlesworth and Willis 2009). This suggests that identification of deleterious alleles may be applied to hybrid breeding strategies (Yang *et al.* 2017). Elevated proportions of deleterious relative to neutral variants in domesticated species suggest a cost of domestication (Moyers *et al.* 2018; Lu *et al.* 2006; Cruz *et al.* 2008; Rodgers-Melnick *et al.* 2015; Liu *et al.* 2017). Studies of the genomic distribution and genetic contribution of deleterious variants can contribute both to understanding the origin and domestication of crop species and to advancing breeding and crop improvement strategies (Morrell *et al.* 2012).

Accurate prediction of deleterious variants is a key component of assessing their contribution to phenotypic variation. Numerous approaches for predicting deleterious variants have been developed. The performance of an approach is typically assessed using the proportion of known, disease-causing human variants that are accurately classified as deleterious. Benchmarking of various approaches using standardized test sets has shown substantial variability among approaches, and improved performance is often achieved through combining results from multiple tools (Thusberg *et al.* 2011; González-Pérez and López-Bigas 2011; Olatubosun *et al.* 2012; Grimm *et al.* 2015). However, the causes of performance differences across approaches are not well understood. While all approaches rely on sequence conservation at the phylogenetic level to identify deleterious variants, some approaches also incorporate protein structure, physical or biochemical properties of amino acid changes, or other attributes of protein sequence when they are available. The earliest conservation metrics used heuristic measures, sometimes including filtering or weighting to account for phylogenetic distance (Sunyaev *et al.* 1999; Miller and Kumar 2001; Ng and Henikoff 2003). More recent approaches have incorporated evolutionary models that account for phylogenetic distance based on putatively neutrally evolving nucleotide sites (Chun and Fay 2009; Davydov *et al.* 2010). Reference bias and the alignments used to calculate conservation metrics are not often emphasized, but are important for making accurate predictions and may account for some of the variability among predictions (Chun and Fay 2009; Hicks *et al.* 2011; Adzhubei *et al.* 2013). Additionally, different predictions have been found using human-based or mouse-based queries of the same substitution (Miosge *et al.* 2015). The accuracy of predictions is particularly dependent on the availability of annotated genomes among related species and the potential to generate sequence alignments.

Despite most approaches being developed for and applied to humans, there has been growing interest in identifying deleterious variants in non-human species in order to understand genomic patterns of variation and their contribution to complex traits, especially in plants. Patterns of deleterious variation have been examined in *Arabidopsis thaliana* (Cao *et al.* 2011), rice (Günther and Schmid 2010; Liu *et al.* 2017), maize (Mezmouk and Ross-Ibarra 2014; Rodgers-Melnick *et al.* 2015), sunflower (Renaut and Rieseberg 2015), poplar (Zhang *et al.* 2016), barley, and soybean (Kono *et al.* 2016). However, the accuracy of predictions in plants has only been examined for a small number of known variants (Günther and Schmid 2010) and only in the past few years have a diverse set of plant genomes and protein homologs become available (Goodstein *et al.* 2012). Furthermore, plants are known to have a larger number of multi-gene families and a higher frequency of polyploidy than occurs in mammals (Lockton and Gaut 2005). These genome-specific factors influence whether a sequence variant is truly deleterious in a given species (Comai 2005; Charlesworth 2012).

The goal of this study was to evaluate the ability of various approaches to predict deleterious variants in plants. The model system *A. thaliana* is a particularly attractive plant species for evaluating approaches that predict deleterious variants because decades of basic research in development, physiology, cell biology, and plant-pathogen interactions have identified large numbers of amino acid-altering mutations with phenotypic consequences. We identified seven approaches that can predict deleterious variants outside of humans (Table S1). Among these approaches, SIFT (Ng and Henikoff 2003), PolyPhen2 (Adzhubei *et al.* 2013) and PROVEAN (Choi *et al.* 2012) generate their own alignments using hits from non-redundant protein databases, whereas MAPP (Stone and Sidow 2005), GERP++ (Davydov *et al.* 2010), and two versions of a likelihood ratio test (Chun and Fay 2009) make predictions using pre-specified alignments as input (Table S1). Because new genome sequences are continually becoming available, the BAD_Mutations pipeline was developed to flexibly identify homologs and generate alignments for any protein of interest (Kono *et al.* 2016). BAD_Mutations uses TBLASTX (Altschul *et al.* 1990) to identify the best match (homolog) from each specified genome and aligns them with PASTA (Mirarab *et al.* 2015). For the four approaches that require alignments, we used the BAD_Mutations pipeline applied to 42 plant genomes. BAD_Mutations was also used to implement two approaches based on a likelihood ratio test (Chun and Fay 2009; Kono *et al.* 2016).

To evaluate predictions of deleterious variants in plants, we generated a curated database of 2,910 *A. thaliana* mutants with known phenotypic alterations. We evaluated the ability of seven approaches to identify these deleterious variants and found that while performance was better than similar assessments in humans, the relative ranking and the highest performing approach differed from previously reported comparisons using human data. Our results demonstrate that reliable prediction of deleterious variants can be achieved in *A. thaliana*, and likely other plant species, expanding the potential value of using deleterious variants to understand naturally occurring variation and to improve crop breeding strategies.

MATERIALS AND METHODS

Generation of a curated set of *Arabidopsis thaliana* mutations

We curated a set of amino acid-altering mutations with phenotypic impacts. Both morphological and biochemical phenotypes were represented, and mutations were in both single-copy and duplicated genes. These mutations were obtained from two sources. We generated a manually curated set of 542 amino acid-altering mutations in 155 genes with phenotypic effects that are described in the literature. These mutations were found by searching the *Arabidopsis* Information Resource (http://www.arabidopsis.org) for genes with either dominant or recessive alleles that differ by nucleotide substitutions. We also identified mutations using a literature search in Google Scholar (http://scholar.google.com). For each variant, we recorded the amino acid substitution, position, and link to the published paper (Table S2). We excluded nonsense mutations because they frequently completely eliminate gene function. We identified a second set of 2,617 amino acid-altering mutations in 960 genes from the manually curated UniProt/Swiss-Prot database (http://www.uniprot.org/) (Boutet *et al.* 2016). The two sets were independently generated and had an overlap of 249 mutants. Using mutants with named alleles as a proxy for those with morphological *vs.* biochemical phenotypes, 65% of our manually curated set and 33% of the Swiss-Prot set had macroscopic phenotypes. Duplicated genes were defined by those proteins with a significant BLASTP hit (E-value < 0.05) to another *A. thaliana* protein with $> 60\%$ identity. By this criterion 466 of 995 proteins were classified as duplicated.

Single nucleotide polymorphisms (SNPs) without any known phenotype were obtained from a set of 80 sequenced *A. thaliana* strains

(Ensembl, version 81, "Cao_SNPs", (Cao *et al.* 2011)). At the time of download, these were the only SNP set available for unrestricted use. After filtering out sites with heterozygous or missing genotype calls, there were 10,797 biallelic amino acid-altering SNPs in the 995 proteins. We used a subset of 1,583 common SNPs (>10%) as those least likely to have phenotypic effects. Our rationale is that on average, strongly deleterious alleles are less likely to reach high frequency in a population, owing to the effects of purifying selection (Fay *et al.* 2001). We also assessed performance by measuring the enrichment of deleterious variants predicted for rare compared to common polymorphisms (Boyko *et al.* 2008). A second set of common amino acid-altering SNPs were identified in an independent set of genes. Excluding the original set of 995 genes, we randomly selected 1,000 proteins from 35,386 peptides in the *A. thaliana* database. We removed 21 that carried no amino acid polymorphism in the 1,001 genomes project (http://www.1001genomes.org). In the remaining 979 genes, we identified 40,736 biallelic amino acid altering SNPs in the 1,001 genomes project, of which 3,717 were common (>10%).

Performance evaluations of seven approaches

We assessed amino acid substitutions using seven approaches: LRT (Chun and Fay 2009), LRT-masked (33), PolyPhen2 (Adzhubei *et al.* 2010), SIFT 4G (Vaser *et al.* 2016), Provean (Choi *et al.* 2012), MAPP (Stone and Sidow 2005) and GERP++ (Davydov *et al.* 2010). PolyPhen2 predictions were generated using the standalone software (v2.2.2) with the PolyPhen2 bundled non-redundant database (uniref100-release 2011_12) and the probabilistic variant classifier using the default HumDiv model. Precomputed SIFT 4G predictions were obtained for *A. thaliana* (TAIR10.23) (http://sift.bii.a-star.edu.sg) and are based on the UniRef90 database (2011). SIFT 4G predictions were not available for 855 substitutions, predominantly because the amino acid change involved more than one nucleotide change within a codon. Provean predictions (v1.1.5) were generated for all mutations using NCBI's non-redundant database (04/02/2016). MAPP and GERP++ predictions were generated using BAD_Mutations alignments and trees (see below). GERP++ generates predictions for single nucleotide positions rather than codons, based on a deficit of observed substitutions compared to that expected given a neutral substitution rate. To assess GERP++ performance we used the GERP++ score at the first, second or third position of the codon if the amino acid substitution could occur by a single change at one of those positions and the average of the GERP++ scores at the first and second positions for all other types of changes. In addition, because GERP++ did not initially perform well on the *A. thaliana* data using neutral substitution rates estimated from each alignment (default) we used a uniform neutral rate of 10 substitutions per site across all genes.

Implementation of BAD_Mutations pipeline

Predictions using a likelihood ratio test (LRT) were performed with the BAD_Mutations pipeline (Kono *et al.* 2016). The pipeline is comprised of Python and Bourne Again Shell (BASH) scripts and incorporates several open-source tools, including the alignment tool PASTA (Mirarab *et al.* 2015) and maximum likelihood methods implemented in HyPhy (Pond *et al.* 2005). The processing step of BAD_Mutations consists of five major subcommands: (1) setup; (2) fetch; (3) align; (4) predict; and (5) compile (Figure S1). The **setup** subcommand generates the configuration files. The **fetch** subcommand downloads gzipped CDS FASTA files from both Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) and Ensembl Plants (http://plants.ensembl.org/index.html), and then creates BLAST databases for identifying homologs. The **align** subcommand uses BLAST to identify homologs of any query protein and generates a protein alignment and phylogenetic tree using PASTA (Mirarab *et al.* 2015). The **predict** subcommand generates predictions for a list of codons of interest by sending a custom batch command to implement a likelihood ratio test using HyPhy. The likelihood ratio test compares the log likelihood of evolution at a single codon under a neutral model ($dN = dS$) to a model allowing for constraint ($dN = \omega dS$), where dN and dS are the synonymous and nonsynonymous substitution rates and ω is a free parameter for selective constraint (Chun and Fay 2009). The **compile** subcommand is to generate the report and p-values. The user manual, including a brief tutorial, is available at https://github.com/MorrellLAB/BAD_Mutations/blob/master/Manual/Manual_v1.0.md.

The BAD_Mutations pipeline makes use of sequenced and annotated genomes. We used BLAST searches of the *A. thaliana* gene sequences against 42 Angiosperm genomes, retaining the top hit from each species with a BLAST E-value threshold of 0.05. The homolog searches were restricted to Angiosperm genomes to avoid extensive saturation of synonymous sites. Protein alignments were generated with PASTA (Mirarab *et al.* 2015), and a likelihood ratio test (LRT) for constraint on each codon of interest was calculated using HyPhy (Pond *et al.* 2005). Sequences with 'N's or other ambiguous nucleotides were discarded prior to the likelihood ratio test. The LRT differs compared to its original formulation (Chun and Fay 2009) in that: i) dS was estimated using all codons for each gene separately, ii) query sequences were optionally masked (the entire sequence changed to N = missing) in the likelihood calculation to avoid any reference bias and iii) branches with dS greater than 3 were set to 3 to avoid spuriously high estimates of dS. Additionally, the original LRT used heuristics to eliminate sites with $dN > dS$, the derived allele present in another species, or sites with fewer than 10 species in the alignment. Rather than eliminating sites, we used logistic regression to provide a single probability of being deleterious based on the LRT test and these additional pieces of information.

Logistic regression was applied using both the masked and unmasked LRT p-values, where the masked p-values were generated from alignments without the *A. thaliana* reference allele. For the unmasked logistic regression, we used the terms log10(LRT p-value), dN/dS, Rn, and An, where Rn and An are the number of *A. thaliana* reference and alternative (*i.e.*, mutant) amino acids observed in the alignment, respectively. For the masked model, we replaced An and Rn with the absolute value of $Rn - An$ and the maximum of Rn and An, respectively. For both models p-values < 1e-16 were set to 1e-16 and constraint values > 10 were set to 10. Ten-fold cross-validation was used to assess the fit of the logistic regression. The average area under the ROC (receiver operating characteristic) curve based on cross-validation was 0.9575 (unmasked) and 0.9471 (masked). Because these values were nearly identical to the performance of the model fit to the entire dataset, 0.9581 (unmasked) and 0.9471 (masked), we used the logistic regression coefficients from the full dataset:

$$\begin{aligned} log(p/(1-p)) = & -2.407 - 0.2139 * LRT(unmasked)) - 0.2056 \\ & * constraint + 0.07368 * Rn - 0.1236 * An \end{aligned}$$

$$\begin{aligned} log(p/(1-p)) = & -2.453 - 0.1904 * LRT(masked) - 0.1459 \\ & * constraint + 0.2199 * max(Rn, An) - 0.2951 \\ & * abs(Rn - An) \end{aligned}$$

Sensitivity, specificity, and area under the curve (AUC) were calculated for each approach using the pROC package in R (Robin *et al.* 2011). We define sensitivity as the proportion of phenotype-altering variants

that are predicted to be deleterious, and specificity as the proportion of variants without known phenotypic effects that are predicted to be neutral. Confidence intervals for each were calculated by 2,000 replicates of stratified bootstrapping, where each replicate contains the same number of positives and negatives as in the original sample.

Combined predictions were generated based on the combined scores of six approaches: LRT, LRT-masked, PolyPhen2, Provean, GERP++, and MAPP. SIFT 4G was not included in the combined predictions because it had missing predictions for a large number (855) variants. Sites with missing predictions from one or more of the remaining approaches (n = 215) were removed. Combined predictions were generated using: 1) logistic regression with each approach's score as a predictive variable, 2) support vector machine, 3) random forest, 4) linear discriminant analysis and 5) generalized linear model with lasso penalized maximum likelihood implemented by the glmnet package in R (Friedman *et al.* 2010). The formula used for the ensemble methods was T∼LRT+LRTm+PPH+PROVEAN+SIFT+GERP+MAPP, where T is a vector of 0/1 for true negative and true positive, and the explanatory terms are the raw scores from each of the singular prediction approaches. The performance of each model was assessed by AUC values obtained from 10-fold cross-validation. The R script is available at https://github.com/MorrellLAB/BAD_Mutations/blob/master/Manuscript_Scripts/script/ensemble.R

Availability of data and materials

LRT predictions were implemented in the Python package BAD_Mutations which is freely available from http://github.com/MorrellLAB/BAD_Mutations.git. All the scripts used for data analysis in this manuscript are available at https://github.com/MorrellLAB/BAD_Mutations/tree/master/Manuscript_Scripts. Alignments of CDS from 42 plant species and Table S2 are available at Data Repository of the University of Minnesota: https://doi.org/10.13020/D6N69S. Supplemental material available at Figshare: https://doi.org/10.25387/g3.6998387.

RESULTS

Curation of a test set of *Arabidopsis thaliana* mutants

To evaluate approaches that predict deleterious variants, we generated a database of *A. thaliana* amino acid substitutions from mutants with described phenotypic alterations and common amino acid polymorphisms unlikely to affect fitness. Out of 2,910 mutants in 995 genes, 81% were from manually curated entries in UniProtKB/Swiss-Prot (n = 2,368), 10% were from our own literature curation (n = 293) and 8.6% were independently identified in both sets (n = 249) (Table S2). Within the same 995 genes, 1,583 common amino acid polymorphisms were identified in 80 accessions (Cao *et al.* 2011). For our analyses, we assume mutations that cause a deviation from the wildtype phenotype are likely deleterious.

Performance of approaches designed to identify deleterious variants

Using the database of *A. thaliana* mutations, we assessed seven approaches for their ability to distinguish deleterious and neutral changes. The approaches were selected because they can generate predictions in non-human organisms. Comparison of sensitivity to specificity showed that each approach could reliably distinguish deleterious and neutral substitutions (Figure 1). A likelihood ratio test (LRT) implemented using the BAD_Mutations pipeline showed significantly higher performance than all other approaches as measured by the area under the curve (AUC) of sensitivity *vs.* specificity (Figure 1, Table S3). A reference masked version of LRT (LRTm), designed to eliminate reference bias (Simons *et al.* 2014), was the approach with the second highest

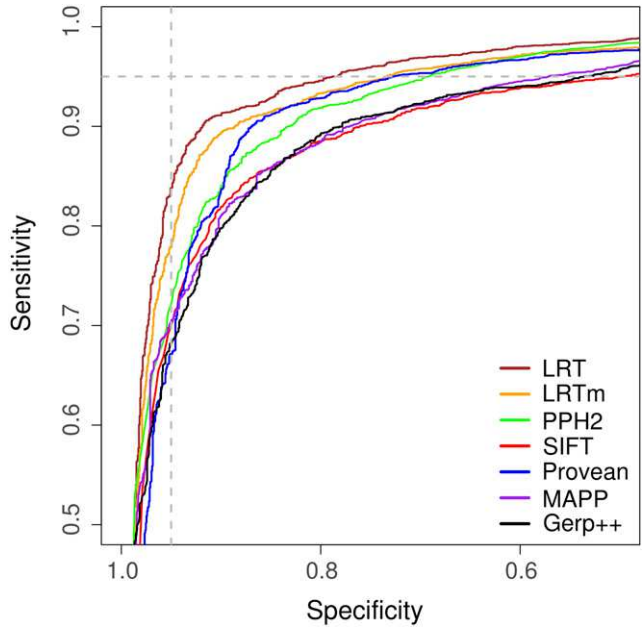

Figure 1 Comparison of approaches that distinguish deleterious and neutral amino acid substitutions. The fraction of true positives (sensitivity) *vs.* the fraction of true negatives (specificity) is shown for seven approaches (LRTm is a masked version of LRT, PPH2 is PolyPhen2). The curves are based on 2,910 deleterious variants and 1,583 neutral variants. Vertical and horizontal dashed lines show the cutoff at 95% specificity and 95% sensitivity, respectively.

performance. PROVEAN and PolyPhen2 showed similar performance as measured by AUC, significantly higher than SIFT, GERP++ and MAPP. The relative ranking by AUC was identical when 1,050 mutations with missing predictions for at least one approach were removed (Table S3). We also found very similar measures of performance when we used common SNPs in a set of independent, randomly selected genes rather than common SNPs within the 995 genes with known phenotype altering mutations (Table S3).

A second means of assessing performance is through comparing predictions of rare *vs.* common variants. Common variants are likely neutral or nearly neutral, whereas deleterious alleles are expected to be kept at low frequency (Ewens 2004). Using SNPs identified in a set of 80 *A. thaliana* strains, we found each approach identified more deleterious SNPs at low compared to common frequencies (Figure 2). At minor allele frequencies between 2/80 (2.5%) and 8/80 (10%), the LRTm and SIFT predicted a lower proportion of deleterious SNPs compared to the other approaches, indicating that they are less sensitive to detecting alleles under weak selection. At the lowest frequency 1/80 (1.25%), which is expected to include many rare and potentially strongly deleterious variants, LRT called the largest proportion of SNPs deleterious.

Performance across phenotypic and duplicate gene categories

To further characterize differences in performance we compared class of variants, including those identified by genome-wide mutant screens or by directly targeting individual proteins. Mutants identified from screens have gross morphological or easily observable phenotypic effects and are often assigned allele names, whereas directed mutants are not often given allele names and tend to have biochemical phenotypes. To compare these two groups, we split the data into those with allele names (1,910), as a proxy for those with gross phenotypes, and those without allele names

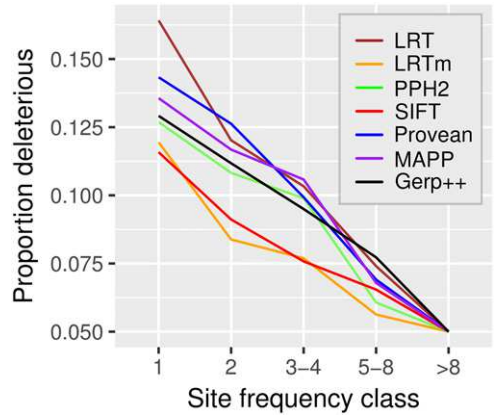

Figure 2 The proportion of SNPs called deleterious across frequency classes. The fraction of SNPs called deleterious by each approach (legend) at its 95% specificity threshold across five frequency classes, labeled by the number of minor alleles present (n = 80). The minor allele is defined as the allele that is less frequent in the sample. Sample sizes for the five classes are 5,303 (1), 1,646 (2), 1,250 (3-4), 1,015 (5-8) and 1,583 (>8).

(1,000), as a proxy for biochemical phenotypes. As measured by AUC, some of the approaches performed better than others and performance was more similar for the gross phenotypic class compared to the biochemical class (Figure 3a). Both SIFT and PolyPhen2 demonstrated the largest difference in performance for predicting mutations with gross phenotypic alterations *vs.* biochemical phenotypes. For this type of mutation, the performance of PolyPhen2 was comparable to the LRT.

Gene duplication may reduce prior selective constraints on a protein, enabling variants to occur at previously conserved sites (Kondrashov *et al.* 2002). Thus, duplicated genes may pose challenges to predicting deleterious alleles, and none of the approaches explicitly distinguish orthologs and paralogs. We identified 466 of the 995 genes as duplicated in *A. thaliana* based on BLASTP hits with 60% or more identity. We compared the performance of these genes to the remaining single copy genes. Each approach showed equal or better performance for duplicated *vs.* single copy genes. SIFT had the largest increase in performance (Figure 3b).

Approach dissimilarity and composite predictions

As reported previously (Doniger *et al.* 2008; Chun and Fay 2009; González-Pérez and López-Bigas 2011; Olatubosun *et al.* 2012), we found substantial disagreement in predictions among the approaches. At a 95% specificity threshold, 93.6% of mutants were predicted deleterious by one or more approach but only 51.3% were predicted deleterious by at least six of the seven approaches (Table S2). Similarly, only 0.25% of common SNPs were predicted deleterious by all approaches but 16.6% were predicted deleterious by at least one approach (LRT and LRTm were considered separately). Comparing the disagreement between approaches, we found LRT and LRTm to produce very similar predictions, but to be distinct from most of the other approaches (Figure 4). We used five models that combined the predictions of all approaches except for SIFT, which had a higher proportion of missing calls. Only two of these ensemble models, a linear discriminant analysis and a generalized linear model with penalized maximum likelihood, performed significantly higher than LRT based on an AUC (Table S4).

DISCUSSION

In this study, we benchmarked the potential for several widely-used approaches to distinguish putatively deleterious and neutral amino acid substitutions in *A. thaliana*. Prior evaluations of performance focused on large sets of mutants for single proteins or known human disease

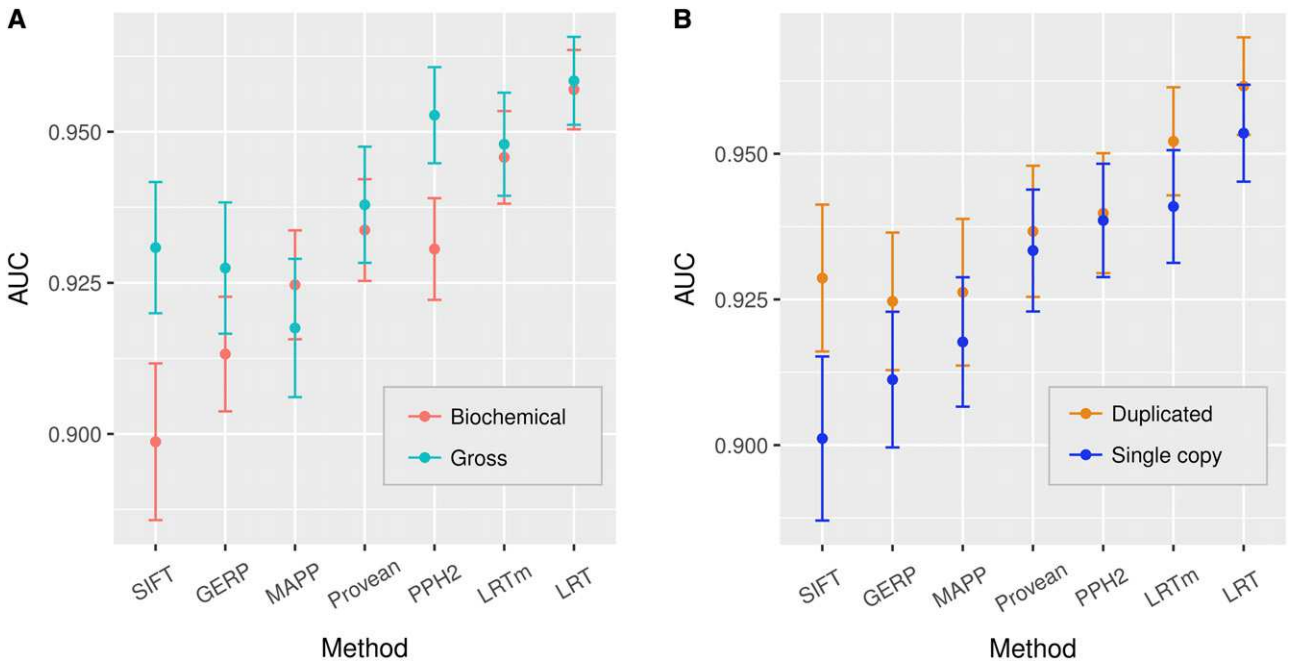

Figure 3 Performance of approaches across different classes of sites. Performance is measured by the area under the curve (AUC) of the approach's sensitivity *vs.* specificity. A – comparison of mutants with biochemical (n = 1,000) *vs.* gross phenotypes (n = 1,910). B – comparison of performance for substitutions in duplicated (n = 2,098) *vs.* single copy genes (n = 2,395). Confidence intervals were determined by 2,000 bootstrapping iterations.

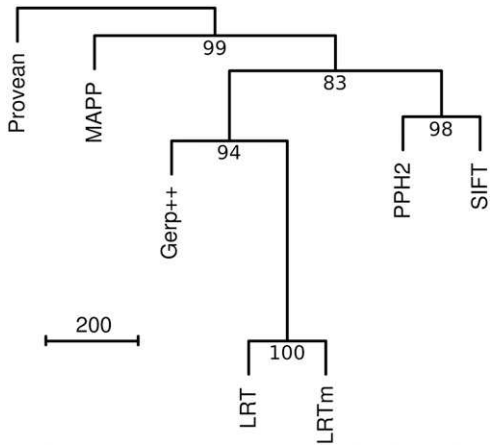

Figure 4 Dissimilarities among approaches. Dissimilarities were computed by the pairwise number of disagreements between each approach applied to mutants and common SNPs (n = 4,493). Dissimilarities are represented by a tree based on hierarchical clustering and values below nodes are bootstrap support based on 2,000 iterations.

variants (Ng and Henikoff 2003; Adzhubei *et al.* 2013). Overall, we find high performance across approaches in their ability to distinguish neutral and deleterious variants, validating their use in plants. The highest performance is achieved by a likelihood ratio test (LRT) implemented using the BAD_Mutations pipeline, in this case using alignments from 42 plant genomes. However, the relative performance depended on the test set and, as discussed below, differs from previous benchmarking studies in humans. Thus, we recommend caution in interpreting slight differences in performance and advocate the use of multiple methods to achieve the highest confidence.

Below, we discuss our results along with characteristics of the approaches and test data that may contribute to differences in predictions and performance when applied to non-human species. One important consideration is the distinction between deleterious variants and those that impact protein function and have phenotypic consequences. While these two groups are overlapping, they are not identical. Because conservation between species is directly related to fitness, we have used the term "deleterious" when referring to the prediction approaches. However, the test sets used to evaluate approaches are composed of variants known to affect protein function or phenotype. Thus, regardless of the nomenclature, any evaluation of approach performance necessarily assumes a large overlap between conserved amino acid positions and those that affect protein function as measured by phenotype. Equally relevant, we use common variants as "neutral" controls even though some common variants are likely to affect protein function due to local adaptation (Hancock *et al.* 2011) or hitchhiking (Chun and Fay 2011). Despite potential contamination, common variants provide the only large set of negative controls that can be used for training and estimating rates of false positives (Ng and Henikoff 2006). Both common and rare variants may also have compensatory effects on deleterious variants (Poon *et al.* 2005). These potential interactions between variants further complicates the identification of truly deleterious variants in any species.

Phylogenetic power, alignments, and reference databases

Phylogenetic power is critical to all comparative genomic approaches that predict deleterious variants. When homologs are too closely related, not enough time has passed for neutral sites to accumulate amino acid substitutions. When homologs are too distantly related, functional sites may not be conserved due to compensatory changes or divergence in homolog function (Marini *et al.* 2010; Breen *et al.* 2012; Jordan *et al.* 2015). The LRT differs from the other approaches examined in that it uses synonymous sites as an internal control to account for the expected amount of protein divergence under a neutral model. As such, even homologs that are nearly identical in their amino acid sequences are informative, given that they have accumulated changes at synonymous sites. However, distantly related homologs are uninformative if divergence at synonymous sites is saturated, thus the LRT should only be applied to organisms where a sufficient number of related genomes are available. In this study, the majority of total dS values for the gene alignments was between 10 and 50, which provides sufficient divergence to test the likelihoods of constraint and relaxation (Chun and Fay 2009). GERP++ is similar to the LRT in that it uses a neutral substitution rate to make its predictions but differs in that the neutral rate must be specified rather than being estimated from synonymous sites within the alignment. GERP++ also does not make use of the genetic code to distinguish synonymous and nonsynonymous changes. In this regard, GERP++ was not appropriately applied since we used a fixed neutral rate for all genes rather than an alignment specific neutral rate.

Out of the approaches compared, phylogenetic power cannot explain the differences between the LRT, MAPP, and GERP++ because they used the same alignments. However, we did notice substantial differences in performance based on the number of ungapped sequences present in the BAD_Mutations alignment at the position being queried (Figure S2). Both LRT and LRTm performed better than the other approaches when there were 10 or fewer sequences at the position of interest. We did not see this pattern when we used the number of sequences present at any position in the alignment, which was typically close to 42. We also did not see this pattern when we examined performance based on the number of sequences used by Provean or PolyPhen2, typically over 100 per gene.

All approaches studied here use alignments to make their predictions, making the protein database and choice of homologs to be included in the alignment a critical step. For MAPP, GERP++, and LRT we used alignments generated using the BAD_Mutations pipeline which queries proteins from a set of annotated reference genomes, in this case from 42 Angiosperm species. SIFT and PolyPhen2 use the UniRef database (2011), whereas PROVEAN uses the most recent non-redundant protein database from NCBI. Both PROVEAN and PolyPhen2 are known to be sensitive to the choice of the reference database and criteria for inclusion of homologs (Choi *et al.* 2012; Adzhubei *et al.* 2013). Despite the choice of homologs being an important step in predicting deleterious substitutions, the use of a plant-specific or entire non-redundant database does not appear to contribute to performance differences: the target database used for prediction does not determine the ranking of approaches in terms of their AUC (Figure 1).

Despite faster runtime of the ensemble approaches with respect to the LRT-based approach, there are circumstances where the LRT-based approach would have higher accuracy. The LRT-based approaches have higher performance in cases where there is shallow alignment depth across the phylogeny, for example, in newly formed genes or rare isoforms of a transcript. The LRT-based approach is able to estimate substitution rates and predict the impact of a variant while the heuristic approaches or the ensemble approaches would likely not make a prediction, and return a missing value.

Training and test sets

Performance of an individual approach depends on both the training and test sets used to measure it. Because performance is typically measured using common SNPs and known disease variants in humans,

Table 1 Performance measured by AUC of approaches based on different test sets

Study	Reference species	Test set	SIFT	PPH2	LRT[1]	GERP++
Dong *et al.* (2015)	Human	SetI	0.76	0.81*	0.72	0.78
	Human	SetII	0.78*	0.76	0.67	0.67
Grimm *et al.* (2015)	Human	VariBenchSelected	0.70*	0.68	0.62	0.59
	Human	predictSNPSelected	0.79	0.79*	0.71	0.67
	Human	SwissVarSelected	0.68	0.71*	0.68	0.65
This study	*A. thaliana*	SwissProt	0.91	0.94	0.96*	0.92
	A. thaliana	Manual curation	0.94	0.96	0.97*	0.94

*Highest performing approach for a given test set.
[1]LRT in this study used a different alignment pipeline than the LRT applied to the human test sets.

there has been some concern over the lack of independence between training and test sets (Dong *et al.* 2015; Grimm *et al.* 2015). However, another consideration that has not yet been examined is whether performance in one species translates to other distantly related species, which may not have the same availability of homologs from sequenced genomes spanning a range of phylogenetic relatedness. The performance of individual approaches could depend on the study system in that some approaches may expect homologs at certain phylogenetic distances, low rates of compensatory change, or low rates of gene duplication.

Previous studies of the accuracy of prediction approaches made use of five human test datasets (Dong *et al.* 2015; Grimm *et al.* 2015). We find better performance across approaches in our *A. thaliana* dataset than that reported for humans (Table 1). It is unclear why the approaches uniformly perform better in *A. thaliana*. One possibility is that the neutral and deleterious variants in *A. thaliana* are more distinct from one another than in humans. The very large proportion of phenotype changing variants in our *A. thaliana* test set that are identified as deleterious means that this test data set is less useful for approach comparison due to the small number of cases that are difficult to predict correctly.

Population and gene-specific performance

Because nearly all measures of performance use either common polymorphism or recently fixed amino acid substitutions as a proxy for neutral SNPs, population and gene-specific factors that influence neutral polymorphism are expected to influence measures of performance. Humans have a small effective population size relative to other mammals (Leffler *et al.* 2012) and consequently a high ratio of nonsynonymous to synonymous diversity (Fay *et al.* 2001; Kosiol *et al.* 2008). Thus, distinguishing neutral and deleterious variants may be more difficult in humans than other species, and approaches trained using human polymorphism may be more conservative with respect to weakly deleterious variants. In comparison, predicting deleterious variants in *A. thaliana* may be facilitated by the fact that *A. thaliana* has slightly larger effective population size (Cao *et al.* 2011).

It should be noted that both demographic history and the process of local adaptation could play important roles in the distribution of deleterious variants. In populations that are colonizing or expanding into novel environments, the selective coefficients against individual variants may change (Slotte *et al.* 2013), and locally adaptive variants may become appreciably enriched. Both humans and *A. thaliana* are known to have undergone demographic expansion in their recent evolutionary histories (Hoffmann 2002; Finlayson 2005). While the relative extent of local adaptation in these two species is difficult to quantify, both exhibit an excess of low-frequency amino acid polymorphism characteristic of deleterious variants (Lohmueller *et al.* 2008; Cao *et al.* 2011; Henn *et al.* 2016).

Another potentially important factor in predicting deleterious variants is gene duplication. *A. thaliana* carries remnants of a whole genome duplication along with numerous tandem duplications (The Arabidopsis Genome Initiative 2000) more than are present in the human genome (Lynch and Conery 2000). Gene duplication can lead to relaxed selection during subfunctionalization or pseudogenization (Ohno 1970), enabling amino acid variants to accumulate in recently duplicated genes. However, we found very similar performance between duplicate and single copy genes, consistent with a similar finding in humans using PolyPhen2 (Adzhubei *et al.* 2013). Because we only included genes with known mutant phenotypes, the sample of recently duplicated genes is limited.

Conclusions and future directions

Most approaches developed to predict deleterious mutations were trained using human data and in many cases, can only be used for human proteins (Li *et al.* 2009; Schwarz *et al.* 2010; Kircher *et al.* 2014). This study demonstrates that several generalized approaches perform exceptionally well in *A. thaliana*, implying that they should also work well for other plant species. Because of the similarly high performance, other considerations such as ease of implementation and compute time may be considered when choosing an approach to identify deleterious mutations in plants. Notably, LRT requires longer run times than any of the other approaches, typically 5.2 hr of computing time per gene compared to 14.5 and 9.4 min per gene for PolyPhen2 and Provean, respectively. One way the BAD_Mutations pipeline could be sped up while retaining the flexibility of querying customizable plant genomes is by using heuristic measures of site-specific conservation rather than the LRT. Provocatively, we found similar performance (AUC = 0.9551) for a logistic model that only used the number of reference and alternative alleles in the alignment (*Rn* and *An*). However, such heuristic measures may not be robust to a change in the reference species and its distance to other genomes in the database. A second approach would be to use predictions from the combined output of multiple prediction approaches, as this has been shown to be highly effective in humans (*e.g.*, González-Pérez and López-Bigas 2011). Although we did not find an ensemble predictor that greatly improved performance, removing LRT predictions did not reduce the performance of the ensemble predictions.

ACKNOWLEDGMENTS

We thank members of the Morrell Lab for discussion and software testing. We also would like to thank Drs. Danelle Seymour and Karl Schmid for helpful comments on an earlier version of the manuscript. Hardware and software support were provided by the University of Minnesota Supercomputing Institute. This work was supported by the US National Science Foundation Plant Genome Program grant (DBI-1339393 to JCF and PLM), the US Department of Agriculture Biotechnology Risk Assessment Research Grants Program (BRAG) (USDA BRAG 2015-06504 to PLM), and a University of Minnesota Doctoral Dissertation Fellowship (to TJYK).

LITERATURE CITED

1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. https://doi.org/10.1038/nature11632

Adzhubei, I., D. M. Jordan, and S. R. Sunyaev, 2013 Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al 0 7: Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.*, 2010 A method and server for predicting damaging missense mutations. Nat. Methods 7: 248–249. https://doi.org/10.1038/nmeth0410-248

Ahituv, N., N. Kavaslar, W. Schackwitz, A. Ustaszewska, J. Martin *et al.*, 2007 Medical sequencing at the extremes of human body mass. Am. J. Hum. Genet. 80: 779–791. https://doi.org/10.1086/513471

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal et al., 2016 UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view, pp. 23–54 in *Plant Bioinformatics*, edited by D. Edwards. Springer, New York, NY.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083. https://doi.org/10.1371/journal.pgen.1000083

Breen, M. S., C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, 2012 Epistasis as the primary factor in molecular evolution. Nature 490: 535–538. https://doi.org/10.1038/nature11510

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43: 956–963. https://doi.org/10.1038/ng.911

Charlesworth, B., 2012 The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. Genetics 191: 233–246. https://doi.org/10.1534/genetics.111.138073

Charlesworth, D., and J. H. Willis, 2009 The genetics of inbreeding depression. Nat. Rev. Genet. 10: 783–796. https://doi.org/10.1038/nrg2664

Choi, Y., G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, 2012 Predicting the functional effect of amino acid substitutions and indels. PLoS One 7: e46688. https://doi.org/10.1371/journal.pone.0046688

Chun, S., and J. C. Fay, 2009 Identification of deleterious mutations within three human genomes. Genome Res. 19: 1553–1561. https://doi.org/10.1101/gr.092619.109

Chun, S., and J. C. Fay, 2011 Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genet. 7: e1002240. https://doi.org/10.1371/journal.pgen.1002240

Comai, L., 2005 The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6: 836–846. https://doi.org/10.1038/nrg1711

Cooper, G. M., and J. Shendure, 2011 Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12: 628–640. https://doi.org/10.1038/nrg3046

Cruz, F., C. Vilà, and M. T. Webster, 2008 The legacy of domestication: accumulation of deleterious mutations in the dog genome. Mol. Biol. Evol. 25: 2331–2336. https://doi.org/10.1093/molbev/msn177

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow *et al.*, 2010 Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLOS Comput. Biol. 6: e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dong, C., P. Wei, X. Jian, R. Gibbs, E. Boerwinkle *et al.*, 2015 Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. 24: 2125–2137. https://doi.org/10.1093/hmg/ddu733

Doniger, S. W., H. S. Kim, D. Swain, D. Corcuera, M. Williams *et al.*, 2008 A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet. 4: e1000183. https://doi.org/10.1371/journal.pgen.1000183

Dudley, J. T., R. Chen, M. Sanderford, A. J. Butte, and S. Kumar, 2012 Evolutionary meta-analysis of association studies reveals ancient constraints affecting disease marker discovery. Mol. Biol. Evol. 29: 2087–2094. https://doi.org/10.1093/molbev/mss079

Ewens, W. J., 2004 *Mathematical population genetics*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21822-9

Fay, J. C., G. J. Wyckoff, and C.-I. Wu, 2001 Positive and negative selection on the human genome. Genetics 158: 1227–1234.

Fay, J. C., 2013 The molecular basis of phenotypic variation in yeast. Curr. Opin. Genet. Dev. 23: 672–677. https://doi.org/10.1016/j.gde.2013.10.005

Finlayson, C., 2005 Biogeography and evolution of the genus *Homo*. Trends Ecol. Evol. 20: 457–463. https://doi.org/10.1016/j.tree.2005.05.019

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33: 1–22. https://doi.org/10.18637/jss.v033.i01

González-Pérez, A., and N. López-Bigas, 2011 Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. Am. J. Hum. Genet. 88: 440–449. https://doi.org/10.1016/j.ajhg.2011.03.004

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes *et al.*, 2012 Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–D1186. https://doi.org/10.1093/nar/gkr944

Grimm, D. G., C.-A. Azencott, F. Aicheler, U. Gieraths, D. G. MacArthur *et al.*, 2015 The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum. Mutat. 36: 513–523. https://doi.org/10.1002/humu.22768

Günther, T., and K. J. Schmid, 2010 Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. Theor. Appl. Genet. 121: 157–168. https://doi.org/10.1007/s00122-010-1299-4

Hancock, A. M., B. Brachi, N. Faure, M. W. Horton, L. B. Jarymowycz *et al.*, 2011 Adaptation to climate across the *Arabidopsis thaliana* genome. Science 334: 83–86. https://doi.org/10.1126/science.1209244

Henn, B. M., L. R. Botigué, S. Peischl, I. Dupanloup, M. Lipatov *et al.*, 2016 Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc. Natl. Acad. Sci. USA 113: E440–E449. https://doi.org/10.1073/pnas.1510805112

Hicks, S., D. A. Wheeler, S. E. Plon, and M. Kimmel, 2011 Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum. Mutat. 32: 661–668. https://doi.org/10.1002/humu.21490

Hoffmann, M. H., 2002 Biogeography of *Arabidopsis thaliana* L. Heynh. (Brassicaceae). J. Biogeogr. 29: 125–134. https://doi.org/10.1046/j.1365-2699.2002.00647.x

Jordan, D. M., S. G. Frangakis, C. Golzio, C. A. Cassa, J. Kurtzberg *et al.*, 2015 Identification of cis-suppression of human disease mutations by comparative genomics. Nature 524: 225–229. https://doi.org/10.1038/nature14497

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper *et al.*, 2014 A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46: 310–315. https://doi.org/10.1038/ng.2892

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, 2002 Selection in the evolution of gene duplications. Genome Biol. 3: research0008.

Kono, T. J. Y., F. Fu, M. Mohammadi, P. J. Hoffman, C. Liu *et al.*, 2016 The role of deleterious substitutions in crop genomes. Mol. Biol. Evol. 33: 2307–2317. https://doi.org/10.1093/molbev/msw102

Kosiol, C., T. Vinař, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante *et al.*, 2008 Patterns of positive selection in six mammalian genomes. PLoS Genet. 4: e1000144. https://doi.org/10.1371/journal.pgen.1000144

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel *et al.*, 2012 Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10: e1001388. https://doi.org/10.1371/journal.pbio.1001388

Li, B., V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati *et al.*, 2009 Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25: 2744–2750. https://doi.org/10.1093/bioinformatics/btp528

Liu, Q., Y. Zhou, P. L. Morrell, and B. S. Gaut, 2017 Deleterious variants in Asian rice and the potential cost of domestication. Mol. Biol. Evol. 34: 908–924. https://doi.org/10.1093/molbev/msw296

Lockton, S., and B. S. Gaut, 2005 Plant conserved non-coding sequences and paralogue evolution. Trends Genet. 21: 60–65. https://doi.org/10.1016/j.tig.2004.11.013

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997. https://doi.org/10.1038/nature06611

Lu, J., T. Tang, H. Tang, J. Huang, S. Shi *et al.*, 2006 The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 22: 126–131. https://doi.org/10.1016/j.tig.2006.01.004

Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753. https://doi.org/10.1038/nature08494

Marini, N. J., P. D. Thomas, and J. Rine, 2010 The use of orthologous sequences to predict the impact of amino acid substitutions on protein function. PLoS Genet. 6: e1000968. https://doi.org/10.1371/journal.pgen.1000968

Mezmouk, S., and J. Ross-Ibarra, 2014 The pattern and distribution of deleterious mutations in maize. G3 (Bethesda)4: 163–171. https://doi.org/10.1534/g3.113.008870

Miller, M. P., and S. Kumar, 2001 Understanding human disease mutations through the use of interspecific genetic variation. Hum. Mol. Genet. 10: 2319–2328. https://doi.org/10.1093/hmg/10.21.2319

Miosge, L. A., M. A. Field, Y. Sontani, V. Cho, S. Johnson *et al.*, 2015 Comparison of predicted and actual consequences of missense mutations. Proc. Natl. Acad. Sci. USA 112: E5189–E5198. https://doi.org/10.1073/pnas.1511585112

Mirarab, S., N. Nguyen, S. Guo, L.-S. Wang, J. Kim *et al.*, 2015 PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol. 22: 377–386. https://doi.org/10.1089/cmb.2014.0156

Morrell, P. L., E. S. Buckler, and J. Ross-Ibarra, 2012 Crop genomics: advances and applications. Nat. Rev. Genet. 13: 85–96. https://doi.org/10.1038/nrg3097

Moyers, B. T., P. L. Morrell, and J. K. McKay, 2018 Genetic costs of domestication and improvement. J. Hered. 109: 103–116. https://doi.org/10.1093/jhered/esx069

Ng, P. C., and S. Henikoff, 2003 SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31: 3812–3814. https://doi.org/10.1093/nar/gkg509

Ng, P. C., and S. Henikoff, 2006 Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. 7: 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630

Ohno, S., 1970 *Evolution by gene duplication*. Springer, Berlin. https://doi.org/10.1007/978-3-642-86659-3

Olatubosun, A., J. Väliaho, J. Härkönen, J. Thusberg, and M. Vihinen, 2012 PON-P: integrated predictor for pathogenicity of missense variants. Hum. Mutat. 33: 1166–1174. https://doi.org/10.1002/humu.22102

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, 2010 Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20: 110–121. https://doi.org/10.1101/gr.097857.109

Pond, S. L. K., S. D. W. Frost, and S. V. Muse, 2005 HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679. https://doi.org/10.1093/bioinformatics/bti079

Poon, A., B. H. Davis, and L. Chao, 2005 The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. Genetics 170: 1323–1332. https://doi.org/10.1534/genetics.104.037259

Renaut, S., and L. H. Rieseberg, 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. https://doi.org/10.1093/molbev/msv106

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek *et al.*, 2011 pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77. https://doi.org/10.1186/1471-2105-12-77

Rodgers-Melnick, E., P. J. Bradbury, R. J. Elshire, J. C. Glaubitz, C. B. Acharya *et al.*, 2015 Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. USA 112: 3823–3828. https://doi.org/10.1073/pnas.1413864112

Schwarz, J. M., C. Rödelsperger, M. Schuelke, and D. Seelow, 2010 MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7: 575–576. https://doi.org/10.1038/nmeth0810-575

Simons, Y. B., M. C. Turchin, J. K. Pritchard, and G. Sella, 2014 The deleterious mutation load is insensitive to recent population history. Nat. Genet. 46: 220–224. https://doi.org/10.1038/ng.2896

Slotte, T., K. M. Hazzouri, J. A. Ågren, D. Koenig, F. Maumus *et al.*, 2013 The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat. Genet. 45: 831–835. https://doi.org/10.1038/ng.2669

Stenson, P. D., M. Mort, E. V. Ball, K. Shaw, A. D. Phillips *et al.*, 2014 The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. 133: 1–9. https://doi.org/10.1007/s00439-013-1358-4

Stone, E. A., and A. Sidow, 2005 Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 15: 978–986. https://doi.org/10.1101/gr.3804205

Sunyaev, S. R., F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan *et al.*, 1999 PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 12: 387–394. https://doi.org/10.1093/protein/12.5.387

The Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815. https://doi.org/10.1038/35048692

Thornton, K. R., A. J. Foran, and A. D. Long, 2013 Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. 9: e1003258. https://doi.org/10.1371/journal.pgen.1003258

Thusberg, J., A. Olatubosun, and M. Vihinen, 2011 Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. 32: 358–368. https://doi.org/10.1002/humu.21445

Vaser, R., S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng, 2016 SIFT missense predictions for genomes. Nat. Protoc. 11: 1–9. https://doi.org/10.1038/nprot.2015.123

Yang, J., S. Mezmouk, A. Baumgarten, E. S. Buckler, K. E. Guill *et al.*, 2017 Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLoS Genet. 13: e1007019. https://doi.org/10.1371/journal.pgen.1007019

Zhang, M., L. Zhou, R. Bawa, H. Suren, and J. A. Holliday, 2016 Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. Mol. Biol. Evol. 33: 2899–2910.

Communicating editor: Y. Kim