Comparative genomics approaches accurately predict deleterious variants in plants

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

[1]*Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN 551085*

2Department of Genetics, Washington University, St. Louis, MO 63110

**Corresponding author: fayjustin@gmail.com & pmorrell@umn.edu*

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 they are nearly always judged 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, the single best-performing approach was a likelihood ratio test applied to homologs identified in 42 plant genomes. Although the approaches did not always agree, we found only slight differences in performance when comparing mutations with gross versus biochemical phenotypes, duplicated versus single copy genes, and when using a single approach versus ensemble predictions. 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 can depend on the organism to which they are applied.

Dramatically increased number of reference genomes, whole genome resequencing, and gene annotations have facilitated the discovery of sequence variants and increased interest in 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). Most approaches assume that variants that disrupt a phylogenetically-conserved site are more likely to be deleterious (Ng and Henikoff 2006). Single nucleotide polymorphisms (SNPs) are the most abundant class of sequence variants. SNPs that alter amino acid sequences are more often associated with phenotypic variation than other classes of variants (1000 Genomes Project Consortium 2012; Fay 2013; Stenson et al. 2014). Amino acid substitutions in protein coding sequences are also the most readily identifiable class of variants that are likely to have biological impact; thus they have been the primary focus of variant annotation efforts.

Annotation of deleterious alleles is also relevant to understanding the genetic basis of phenotypic variation in other species. Complementation of recessive 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. 2016). Annotation of deleterious variants improves prediction accuracy of complex traits (Dudley et al. 2012). Elevated proportions of deleterious relative to neutral variants in domesticated species suggest a cost of domestication (Cruz et al. 2008; Liu et al. 2017; Lu et al. 2006; Rodgers-Melnick et al. 2015). 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 (Ng and Henikoff 2006). The performance of an approach is typically assessed using the proportion of known, human disease causing variants that are accurately classified as deleterious. Benchmarking of various approaches using uniform test sets has shown substantial variability among approaches and improved performance is often achieved through the use of ensemble predictions based on multiple predictions (González-Pérez and López-Bigas 2011; Grimm et al. 2015; Olatubosun et al. 2012; Thusberg et al. 2011). 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 (Ng and Henikoff 2003). More recent approaches have incorporated evolutionary models that account for phylogenetic distance based on putatively neutrally evolving nucleotide sites (Davydov et al. 2010; Chun and Fay 2009). Reference bias and the alignments used to calculate conservation metrics are not often emphasized, but are important to accurate predictions and may account for some of the variability among predictions (Adzhubei et al. 2013; Chun and Fay 2009; Hicks et al. 2011). The accuracy of predictions is particularly dependent on the availability of annotated genomes among related species and the potential to generate sequence alignments, particularly for protein coding regions of the genome.

Studies of deleterious variants in non-human species are limited to a small subset of approaches that are not human-specific. Even so, there is a growing body of research that uses predicted deleterious variants 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 (Charlesworth 2012). 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.

To evaluate various tools to predict deleterious variants in plants, we generated a curated database of 2,910 *A. thaliana* mutants with known phenotypic alterations. We identified seven approaches that can predict deleterious variants outside of humans. Among these approaches, SIFT (Ng and Henikoff 2003), PolyPhen2 (Adzhubei et al. 2013) and PROVEAN (Choi et al. 2012) generate their own alignments using 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. For these latter cases we used the BAD_Mutations pipeline for identifying homologs and alignments based on 42 plant genomes (Kono et al. 2016). We found that while all approaches performed better than similar assessments in humans, the relative ranking and the highest performing approach differed from previously reported comparisons using human data. We did not find factors that are major determinants of differences among approaches. 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 better understand naturally occurring variation and to improve crop breeding strategies.

Results

A database of literature curated *Arabidopsis thaliana* mutants

To evaluate approaches that predict deleterious variants, we generated a database of *A. thaliana* amino acid substitutions from *i*) mutants with described phenotypic alterations and *ii*) 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 S1). 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 versus specificity as well as at thresholds of 95% sensitivity and 95% specificity (Figure 1, Table S1). A reference masked version of LRT (LRTm), designed to eliminate reference bias (Simons et al. 2014), was the approach with the second highest 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 S1).

A second means of assessing performance is through comparing predictions of rare versus common variants. Common variants are likely neutral or nearly neutral, whereas deleterious alleles are 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. In general, mutants identified from screens have gross morphological or easily observable phenotypic effects and are assigned allele names, whereas directed mutants are typically not given allele names and 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 (1,000), as a proxy for biochemical phenotypes. As measured by AUC, some of the approaches performed better and their performance was more similar for the gross phenotypic class compared to the biochemical class (Figure 3a). Both SIFT and PolyPhen2 demonstrated the largest increase in performance for predicting mutations with gross phenotypic alterations. 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 versus single copy genes, with SIFT in particular showing the largest increase in performance (Figure 3b).

Approach dissimilarity and composite predictions

A reported previously (Chun and Fay 2009; Doniger et al. 2008; 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 six or more of the seven approaches. Similarly, only 0.25% of common SNPs were predicted deleterious by all approaches but 16.6% were predicted deleterious by at least one. 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 S2).

Discussion

In this study, we benchmarked the ability of 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 variants (Adzhubei et al.

2013; Ng and Henikoff 2003). 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. Despite considerable variation among prediction approaches, no single factor explains differences among performance.

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 and divergence 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.

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 (Breen et al. 2012; Jordan et al. 2015; Marini et al. 2010). 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, so long as 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. 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++ since they used the same alignments.

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 sequenced plant 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 (Adzhubei et al. 2013; Choi et al. 2012) . 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 be a major contributor to performance differences (Figure 1).

Training and test sets

Performance of an individual approaches 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, 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 phenotyping 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 it is a selfing species with an effective population size larger than that of humans (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. 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; Henn et al. 2016; Cao et al. 2011).

Another potentially important factor in predicting deleterious variants is gene duplication. *A. thaliana* carries remnants of a whole genome duplication along with numerous single copy 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. Recent duplicates are more likely to accumulate common variants that would appear deleterious and so may be among those genes where predictions are the most difficult.

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, e.g., Kircher et al. 2014; Li et al. 2009; Schwarz et al. 2010. This study demonstrates that several generalized approaches perform exceptionally well in *A. thaliana*, implying that they should also work well for other plant species. Despite the high performance, it is quite likely further improvements could be achieved. Notably, LRT requires longer run times than any of the other approaches, typically 5.2 hrs of compute time per gene. Although we did not investigate whether a faster approach could be implemented without a loss in performance, it is acknowledged that the long run time of the LRT may limit the application of the approach to large genomics datasets. One potential avenue to pursue is whether faster heuristic measures of site-specific conservation based on the BAD_Mutations pipeline of alignments could achieve similarly high performance. However, further study would be needed to test whether heuristic measures of amino acid conservation would be robust to the reference species and protein alignments to which they were applied. A second approach would be to find a more effective means of generating 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, this might reflect the relatively small number of variables used to generate ensemble predictions.

Methods

Mutations with phenotypic effects 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 caused 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 S3). 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 those mutants with named alleles as an indicator of those with gross versus 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 greater than 60% identity. By this criteria 466/995 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 with 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.

We assessed amino acid substitutions using six approaches: LRT (Chun and Fay 2009), 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 predictions were generated using BAD_Mutations alignments and trees (see below). GERP++ generates predictions for single nucleotide positions rather than codons. 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 perform well using neutral substitution rates estimated from each alignment (default) we used a uniform neutral rate of 10 substitutions per site across all genes.

Predictions using a likelihood ratio test (LRT) were performed with the BAD_Mutations pipeline (Kono et al. 2016). The pipeline makes use of sequenced and annotated genomes. We used blast searches of 42 angiosperm genomes and retaining the top hit from each with a blast e-value threshold of 0.05. Only Angiosperms were used to avoid extensive saturation of synonymous sites. Pasta protein alignments (Mirarab et al. 2015) were generated using the homologs and the likelihood of $dN = \omega dS$ compared to $dN = dS$ for each codon of interest was calculated using HYPHY (Pond et al. 2005), where dN and dS are the nonsynonymous and synonymous substitution rate and ω is a free parameter. Sequences with 'N's or other ambiguous nucleotides were discarded prior to the likelihood ratio test. The LRT differs compared to its original formation (Chun and Fay 2009) in that: i) dS was estimated using all codons for each gene separately, ii) query sequences were optionally masked 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 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), constraint (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 less than 1e-16 were set to 1e-16 and constraint values greater than 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 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:

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

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

Sensitivity, specificity and area under the curve (AUC) were calculated for each approach using the pROC package in R (Robin et al. 2011). Confidence intervals for each were calculated by stratified bootstrapping (*n* = 2000).

Combined predictions were generated based on the combined scores of six approaches: LRT, LRT-masked, PolyPhen2, Provean, GERP++ and MAPP. Sites with missing predictions from one or more approach (*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 penalized maximum likelihood implemented by the glmnet package in R (Friedman et al. 2010). The performance of each model was assessed by AUC values obtained from 10-fold cross-validation.

Data access

LRT predictions were implemented in the Python package BAD_Mutations which is freely available from http://github.com/MorrellLAB/BAD_Mutations.git.

Acknowledgements

We acknowledge funding support from the US National Science Foundation Plant Genome Program grant (DBI-1339393 to JCF and PLM), from a University of Minnesota Doctoral Dissertation Fellowship supporting TJYK. We thank members of the Morrell Lab for discussion and software testing. Finally, we would like to thank Dr. Danelle Seymour from UC Irvine for comments.

Author contributions:

T.J.Y.K. and P.J.H. wrote code for BAD_Mutations. J.C.F., L.L. and C.H. S. analyzed the data. L.L., J.C.F., T.J.Y.K., and P.L.M. wrote the initial draft of the manuscript. All authors contributed to final manuscript preparation.

Reference:

1000 Genomes Project Consortium T 1000 GP. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet Editor Board Jonathan Haines Al* **0 7**: Unit7.20.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248–249.

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hébert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, et al. 2007. Medical sequencing at the extremes of human body mass. *Am J Hum Genet* **80**: 779–791.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In *Plant Bioinformatics* (ed. D. Edwards), Vol. 1374 of, pp. 23–54, Springer New York, New York, NY http://link.springer.com/10.1007/978-1-4939-3167-5_2 (Accessed January 19, 2017).

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. *Nature* **490**: 535–538.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**: 956–963.

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.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. *Nat Rev Genet* **10**: 783–796.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. *PLOS ONE* **7**: e46688.

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

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**: 628–640.

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.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput Biol* **6**: e1001025.

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. 2015. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**: 2125–2137.

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

Dudley JT, Chen R, Sanderford M, Butte AJ, Kumar S. 2012. Evolutionary meta-analysis of association studies reveals ancient constraints affecting disease marker discovery. *Mol Biol Evol* **29**: 2087–2094.

Ewens WJ. 2004. *Mathematical population genetics*. Springer New York, New York, NY http://link.springer.com/10.1007/978-0-387-21822-9 (Accessed January 19, 2017).

Fay JC. 2013. The molecular basis of phenotypic variation in yeast. *Curr Opin Genet Dev* **23**: 672–677.

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

Finlayson C. 2005. Biogeography and evolution of the genus *Homo*. *Trends Ecol Evol* **20**: 457–463.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**: 1–22.

González-Pérez A, López-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. *Am J Hum Genet* **88**: 440–449.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–D1186.

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, 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.

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

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci* **113**: E440–E449.

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum Mutat* **32**: 661–668.

Hoffmann MH. 2002. Biogeography of *Arabidopsis thaliana* L. Heynh. (Brassicaceae). *J Biogeogr* **29**: 125–134.

Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal Genomics, Davis EE, Sunyaev SR, Katsanis N. 2015. Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* **524**: 225–229.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310–315.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3**: research0008.

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.

Kosiol C, Vinař T, Fonseca RR da, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLOS Genet* **4**: e1000144.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLOS Biol* **10**: e1001388.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**: 2744–2750.

Liu Q, Zhou Y, Morrell PL, Gaut BS. 2017. Deleterious variants in Asian rice and the potential cost of domestication. *Mol Biol Evol* msw296.

Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet* **21**: 60–65.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**: 994–997.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet* **22**: 126–131.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

Marini NJ, Thomas PD, Rine J. 2010. The use of orthologous sequences to predict the impact of amino acid substitutions on protein function. *PLOS Genet* **6**: e1000968.

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

Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J Comput Biol* **22**: 377–386.

Morrell PL, Buckler ES, Ross-Ibarra J. 2012. Crop genomics: advances and applications. *Nat Rev Genet* **13**: 85–96.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**: 61–80.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**: 3812–3814.

Ohno S. 1970. *Evolution by gene duplication*. Springer Berlin Heidelberg, Berlin, Heidelberg http://link.springer.com/10.1007/978-3-642-86659-3 (Accessed January 19, 2017).

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat* **33**: 1166–1174.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

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.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**: 77.

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* **112**: 3823–3828.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**: 575–576.

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

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45**: 831–835.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. 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.

Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* **15**: 978–986.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

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.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* **32**: 358–368.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc* **11**: 1–9.

Yang J, Mezmouk S, Baumgarten A, Buckler ES, Guill KE, McMullen MD, Mumm RH, Ross-Ibarra J. 2016. Incomplete dominance of deleterious alleles contribute substantially to trait variation and heterosis in maize. *bioRxiv* 86132.

Zhang M, Zhou L, Bawa R, Suren H, Holliday JA. 2016. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Mol Biol Evol* msw169.

Tables

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
Grim 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.

Figure Legends

Figure 1. Comparison of approaches that distinguish deleterious and neutral amino acid substitutions. The fraction of true positives (sensitivity) versus the fraction of true negatives (specificity) is shown for seven approaches (LRTm is a masked version of LRT, PPH2 is PolyPhen2). Vertical and horizontal

dashed lines show the cutoff at 95% specificity and 95% sensitivity, respectively.

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). Sample sizes for the five classes are 5303 (1), 1646 (2), 1250 (3-4), 1015 (5-8) and 1583 (>8).

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 versus specificity. A – comparison of mutants with biochemical versus gross phenotypes. B – comparison of performance for duplicated versus single copy genes.

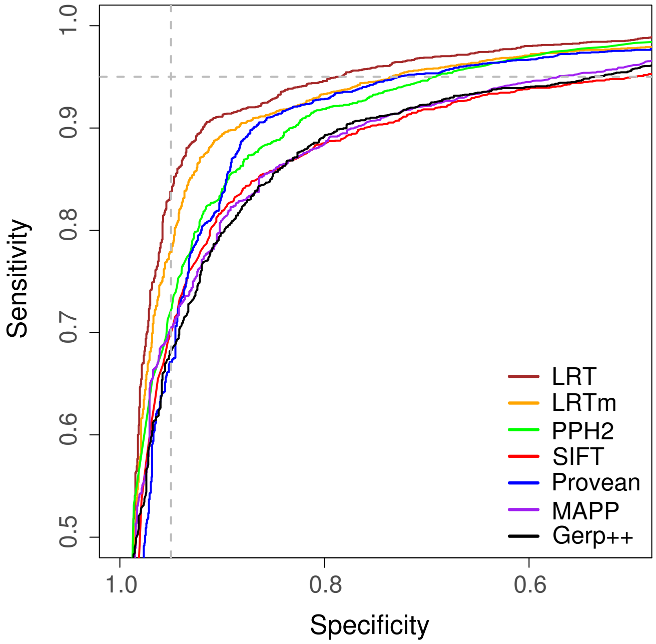
Figure 4. Dissimilarities among approaches. Dissimilarities were computed by the pairwise number of disagreements between each approach applied to mutants and common SNPs. Dissimilarities are represented by a tree based on hierarchical clustering.

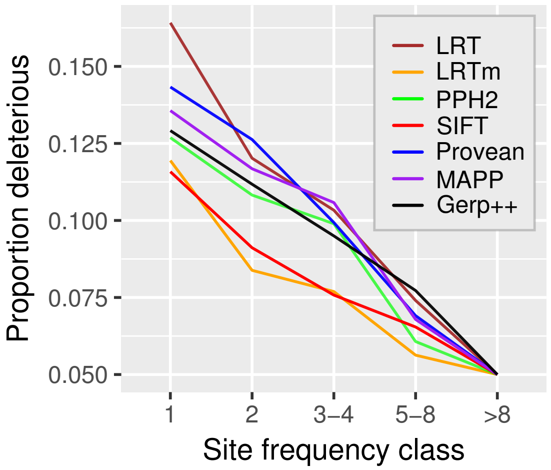
Supplemental Tables

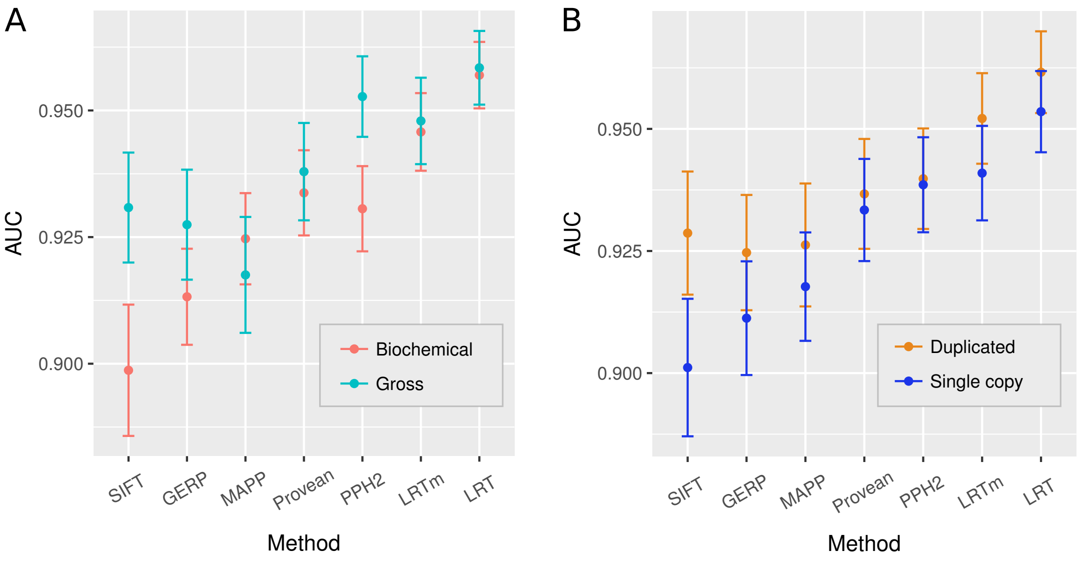
Table S1. Performance of methods used to distinguish deleterious and neutral substitutions

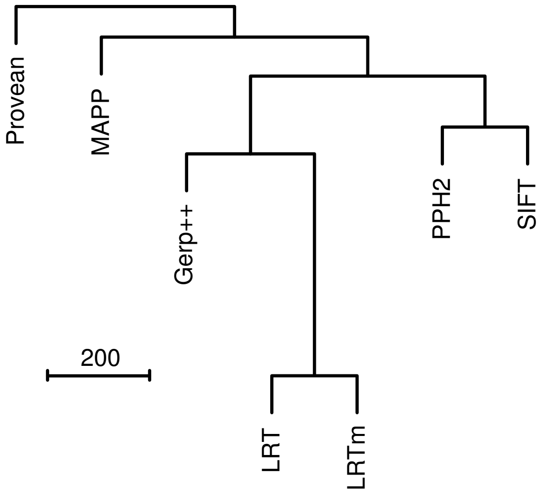
Table S2. Performance of models based on ensemble prediction methods.

Table S3. 2,617 amino acid altering mutations in 960 *A. thaliana* genes


Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5
Specificity
1.0
0.8
0.6
LRT
LRTm
PPH2
SIFT
Provean
MAPP
Gerp++


Proportion deleterious
0.150
0.125
0.100
0.075
0.050
1
2
3–4
5–8
>8
Site frequency class
LRT
LRTm
PPH2
SIFT
Provean
MAPP
Gerp++


A
AUC
0.950
0.925
0.900
SIFT
GERP
MAPP
Provean
PPH2
LRTm
LRT
Method
Biochemical
Gross
B
AUC
0.950
0.925
0.900
SIFT
GERP
MAPP
Provean
PPH2
LRTm
LRT
Method
Duplicated
Single copy


Provean
MAPP
Gerp++
LRT
LRTm
PPH2
SIFT
200