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*

2*Department 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

24 versus single copy genes, and when using a single approach versus ensemble predictions. We
25 conclude that deleterious mutations can be reliably predicted in *A. thaliana* and likely other plant
26 species, but that the relative performance of various approaches can depend on the organism to which
27 they are applied.
28
29
30 Dramatically increased number of reference genomes, whole genome resequencing, and gene
31 annotations have facilitated the discovery of sequence variants and increased interest in annotation of
32 functional variants in many organisms. Functional annotation can yield insight into the genetic basis
33 of phenotypic variation and is often a critical step in the identification of genes and variants
34 underlying human disease (Ahituv et al. 2007; Cooper and Shendure 2011). In particular, interest in
35 identifying putatively deleterious variants has increased, because these variants may contribute
36 substantially to phenotypic variation (Manolio et al. 2009; Thornton et al. 2013). Most approaches
37 assume that variants that disrupt a phylogenetically-conserved site are more likely to be deleterious
38 (Ng and Henikoff 2006). Single nucleotide polymorphisms (SNPs) are the most abundant class of
39 sequence variants. SNPs that alter amino acid sequences are more often associated with phenotypic
40 variation than other classes of variants (1000 Genomes Project Consortium 2012; Fay 2013; Stenson et
41 al. 2014). Amino acid substitutions in protein coding sequences are also the most readily identifiable
42 class of variants that are likely to have biological impact; thus they have been the primary focus of
43 variant annotation efforts.
44 Annotation of deleterious alleles is also relevant to understanding the genetic basis of
45 phenotypic variation in other species. Complementation of recessive deleterious variants between

46 haplotypes is thought to be one of the primary mechanisms underlying heterosis (Charlesworth and
47 Willis 2009). This suggests that identification of deleterious alleles may be applied to hybrid breeding
48 strategies (Yang et al. 2016). Annotation of deleterious variants improves prediction accuracy of
49 complex traits (Dudley et al. 2012). Elevated proportions of deleterious relative to neutral variants in
50 domesticated species suggest a cost of domestication (Cruz et al. 2008; Liu et al. 2017; Lu et al. 2006;
51 Rodgers-Melnick et al. 2015). Studies of the genomic distribution and genetic contribution of
52 deleterious variants can contribute both to understanding the origin and domestication of crop species
53 and to advancing breeding and crop improvement strategies (Morrell et al. 2012).

54 Accurate prediction of deleterious variants is a key component of assessing their contribution
55 to phenotypic variation. Numerous approaches for predicting deleterious variants have been
56 developed (Ng and Henikoff 2006). The performance of an approach is typically assessed using the
57 proportion of known, human disease causing variants that are accurately classified as deleterious.
58 Benchmarking of various approaches using uniform test sets has shown substantial variability among
59 approaches and improved performance is often achieved through the use of ensemble predictions
60 based on multiple predictions (González-Pérez and López-Bigas 2011; Grimm et al. 2015; Olatubosun
61 et al. 2012; Thusberg et al. 2011). However, the causes of performance differences across approaches
62 are not well understood. While all approaches rely on sequence conservation at the phylogenetic level
63 to identify deleterious variants, some approaches also incorporate protein structure, physical or
64 biochemical properties of amino acid changes, or other attributes of protein sequence when they are
65 available. The earliest conservation metrics used heuristic measures, sometimes including filtering or
66 weighting to account for phylogenetic distance (Ng and Henikoff 2003). More recent approaches have
67 incorporated evolutionary models that account for phylogenetic distance based on putatively

68 neutrally evolving nucleotide sites (Davydov et al. 2010; Chun and Fay 2009). Reference bias and the
69 alignments used to calculate conservation metrics are not often emphasized, but are important to
70 accurate predictions and may account for some of the variability among predictions (Adzhubei et al.
71 2013; Chun and Fay 2009; Hicks et al. 2011). The accuracy of predictions is particularly dependent on
72 the availability of annotated genomes among related species and the potential to generate sequence
73 alignments, particularly for protein coding regions of the genome.

74 Studies of deleterious variants in non-human species are limited to a small subset of
75 approaches that are not human-specific. Even so, there is a growing body of research that uses
76 predicted deleterious variants to understand genomic patterns of variation and their contribution to
77 complex traits, especially in plants. Patterns of deleterious variation have been examined in
78 *Arabidopsis thaliana (Cao et al. 2011)*, rice (Günther and Schmid 2010; Liu et al. 2017), maize (Mezmouk
79 and Ross-Ibarra 2014; Rodgers-Melnick et al. 2015), sunflower (Renaut and Rieseberg 2015), poplar
80 (Zhang et al. 2016), barley, and soybean (Kono et al. 2016). However, the accuracy of predictions in
81 plants has only been examined for a small number of known variants (Günther and Schmid 2010) and
82 only in the past few years have a diverse set of plant genomes and protein homologs become available
83 (Goodstein et al. 2012). Furthermore, plants are known to have a larger number of multi-gene families
84 and a higher frequency of polyploidy than occurs in mammals (Lockton and Gaut 2005). These
85 genome-specific factors influence whether a sequence variant is truly deleterious (Charlesworth 2012).
86 The model system *A. thaliana* is a particularly attractive plant species for evaluating approaches that
87 predict deleterious variants because decades of basic research in development, physiology, cell
88 biology, and plant-pathogen interactions have identified large numbers of amino acid altering
89 mutations with phenotypic consequences.

90 To evaluate various tools to predict deleterious variants in plants, we generated a curated
91 database of 2,910 *A. thaliana* mutants with known phenotypic alterations. We identified seven
92 approaches that can predict deleterious variants outside of humans. Among these approaches, SIFT
93 (Ng and Henikoff 2003), PolyPhen2 (Adzhubei et al. 2013) and PROVEAN (Choi et al. 2012) generate
94 their own alignments using non-redundant protein databases, whereas MAPP (Stone and Sidow
95 2005), GERP++ (Davydov et al. 2010), and two versions of a likelihood ratio test (Chun and Fay 2009)
96 make predictions using pre-specified alignments as input. For these latter cases we used the
97 BAD_Mutations pipeline for identifying homologs and alignments based on 42 plant genomes (Kono
98 et al. 2016). We found that while all approaches performed better than similar assessments in humans,
99 the relative ranking and the highest performing approach differed from previously reported
100 comparisons using human data. We did not find factors that are major determinants of differences
101 among approaches. Our results demonstrate that reliable prediction of deleterious variants can be
102 achieved in *A. thaliana* and likely other plant species, expanding the potential value of using
103 deleterious variants to better understand naturally occurring variation and to improve crop breeding
104 strategies.
105

106 Results

107

108 A database of literature curated *Arabidopsis thaliana* mutants

109 To evaluate approaches that predict deleterious variants, we generated a database of *A. thaliana* amino
110 acid substitutions from *i*) mutants with described phenotypic alterations and *ii*) common amino acid
111 polymorphisms unlikely to affect fitness. Out of 2,910 mutants in 995 genes, 81% were from manually
112 curated entries in UniProtKB/Swiss-Prot (n = 2,368), 10% were from our own literature curation (n =

113 293) and 8.6% were independently identified in both sets (n = 249) (Table S1). Within the same 995
114 genes, 1,583 common amino acid polymorphisms were identified in 80 accessions (Cao et al. 2011). For
115 our analyses, we assume mutations that cause a deviation from the wildtype phenotype are likely
116 deleterious.
117

118 Performance of approaches designed to identify deleterious variants

119 Using the database of *A. thaliana* mutations, we assessed seven approaches for their ability to
120 distinguish deleterious and neutral changes. The approaches were selected because they can generate
121 predictions in non-human organisms. Comparison of sensitivity to specificity showed that each
122 approach could reliably distinguish deleterious and neutral substitutions (Figure 1). A likelihood ratio
123 test (LRT) implemented using the BAD_Mutations pipeline showed significantly higher performance
124 than all other approaches as measured by the area under the curve (AUC) of sensitivity versus
125 specificity as well as at thresholds of 95% sensitivity and 95% specificity (Figure 1, Table S1). A
126 reference masked version of LRT (LRTm), designed to eliminate reference bias (Simons et al. 2014),
127 was the approach with the second highest performance. PROVEAN and PolyPhen2 showed similar
128 performance as measured by AUC, significantly higher than SIFT, GERP++ and MAPP. The relative
129 ranking by AUC was identical when 1,050 mutations with missing predictions for at least one
130 approach were removed (Table S1).
131 A second means of assessing performance is through comparing predictions of rare versus
132 common variants. Common variants are likely neutral or nearly neutral, whereas deleterious alleles
133 are kept at low frequency (Ewens 2004). Using SNPs identified in a set of 80 *A. thaliana* strains, we
134 found each approach identified more deleterious SNPs at low compared to common frequencies

135 (Figure 2). At minor allele frequencies between 2/80 (2.5%) and 8/80 (10%) the LRTm and SIFT
136 predicted a lower proportion of deleterious SNPs compared to the other approaches, indicating that
137 they are less sensitive to detecting alleles under weak selection. At the lowest frequency 1/80 (1.25%),
138 which is expected to include many rare and potentially strongly deleterious variants, LRT called the
139 largest proportion of SNPs deleterious.
140

141 Performance across phenotypic and duplicate gene 142 categories

143 To further characterize differences in performance we compared class of variants, including those
144 identified by genome-wide mutant screens or by directly targeting individual proteins. In general,
145 mutants identified from screens have gross morphological or easily observable phenotypic effects and
146 are assigned allele names, whereas directed mutants are typically not given allele names and have
147 biochemical phenotypes. To compare these two groups, we split the data into those with allele names
148 (1,910), as a proxy for those with gross phenotypes, and those without allele names (1,000), as a proxy
149 for biochemical phenotypes. As measured by AUC, some of the approaches performed better and
150 their performance was more similar for the gross phenotypic class compared to the biochemical class
151 (Figure 3a). Both SIFT and PolyPhen2 demonstrated the largest increase in performance for predicting
152 mutations with gross phenotypic alterations. For this type of mutation, the performance of PolyPhen2
153 was comparable to the LRT.

154 Gene duplication may reduce prior selective constraints on a protein, enabling variants to
155 occur at previously conserved sites (Kondrashov et al. 2002). Thus duplicated genes may pose
156 challenges to predicting deleterious alleles and none of the approaches explicitly distinguish orthologs

157 and paralogs. We identified 466 of the 995 genes as duplicated in *A. thaliana* based on blastp hits with
158 60% or more identity. We compared the performance of these genes to the remaining single copy
159 genes. Each approach showed equal or better performance for duplicated versus single copy genes,
160 with SIFT in particular showing the largest increase in performance (Figure 3b).
161

162 Approach dissimilarity and composite predictions

163 A reported previously (Chun and Fay 2009; Doniger et al. 2008; González-Pérez and López-Bigas 2011;
164 Olatubosun et al. 2012), we found substantial disagreement in predictions among the approaches. At a
165 95% specificity threshold, 93.6% of mutants were predicted deleterious by one or more approach but
166 only 51.3% were predicted deleterious by six or more of the seven approaches. Similarly, only 0.25%
167 of common SNPs were predicted deleterious by all approaches but 16.6% were predicted deleterious
168 by at least one. Comparing the disagreement between approaches we found LRT and LRTm to
169 produce very similar predictions, but to be distinct from most of the other approaches (Figure 4). We
170 used five models that combined the predictions of all approaches except for SIFT, which had a higher
171 proportion of missing calls. Only two of these ensemble models, a linear discriminant analysis and a
172 generalized linear model with penalized maximum likelihood, performed significantly higher than
173 LRT based on an AUC (Table S2).
174

175 Discussion

176

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

180 2013; Ng and Henikoff 2003). Overall we find high performance across approaches in their ability to
181 distinguish neutral and deleterious variants, validating their use in plants. The highest performance is
182 achieved by a likelihood ratio test (LRT) implemented using the BAD_Mutations pipeline, in this case
183 using alignments from 42 plant genomes. Despite considerable variation among prediction
184 approaches, no single factor explains differences among performance.

185 Below, we discuss our results along with characteristics of the approaches and test data that
186 may contribute to differences in predictions and performance when applied to non-human species.
187 One important consideration is the distinction between deleterious variants and those that impact
188 protein function and have phenotypic consequences. While these two groups are overlapping, they
189 are not identical. Because conservation and divergence between species is directly related to fitness,
190 we have used the term "deleterious" when referring to the prediction approaches. However, the test
191 sets used to evaluate approaches are composed of variants known to affect protein function or
192 phenotype. Thus, regardless of the nomenclature any evaluation of approach performance necessarily
193 assumes a large overlap between conserved amino acid positions and those that affect protein
194 function as measured by phenotype.

195

196 Phylogenetic power, alignments, and reference databases

197 Phylogenetic power is critical to all comparative genomic approaches that predict deleterious variants.
198 When homologs are too closely related, not enough time has passed for neutral sites to accumulate
199 amino acid substitutions. When homologs are too distantly related, functional sites may not be
200 conserved due to compensatory changes or divergence in homolog function (Breen et al. 2012; Jordan
201 et al. 2015; Marini et al. 2010). The LRT differs from the other approaches examined in that it uses

202 synonymous sites as an internal control to account for the expected amount of protein divergence
203 under a neutral model. As such, even homologs that are nearly identical in their amino acid sequences
204 are informative, so long as they have accumulated changes at synonymous sites. However, distantly
205 related homologs are uninformative if divergence at synonymous sites is saturated, thus the LRT
206 should only be applied to organisms where a sufficient number of related genomes are available.
207 GERP++ is similar to the LRT in that it uses a neutral substitution rate to make its predictions, but
208 differs in that the neutral rate must be specified rather than being estimated from synonymous sites
209 within the alignment. GERP++ also does not make use of the genetic code to distinguish synonymous
210 and nonsynonymous changes. In this regard, GERP++ was not appropriately applied since we used a
211 fixed neutral rate for all genes rather than an alignment specific neutral rate. Out of the approaches
212 compared, phylogenetic power cannot explain the differences between the LRT, MAPP and GERP++
213 since they used the same alignments.

214 All approaches studied here use alignments to make their predictions, making the protein
215 database and choice of homologs to be included in the alignment a critical step. For MAPP, GERP++,
216 and LRT we used alignments generated using the BAD_Mutations pipeline which queries proteins
217 from sequenced plant genomes, in this case from 42 Angiosperm species. SIFT and PolyPhen2 use the
218 UniRef database (2011), whereas PROVEAN uses the most recent non-redundant protein database
219 from NCBI. Both PROVEAN and PolyPhen2 are known to be sensitive to the choice of the reference
220 database and criteria for inclusion of homologs (Adzhubei et al. 2013; Choi et al. 2012) . Despite the
221 choice of homologs being an important step in predicting deleterious substitutions, the use of a plant-
222 specific or entire non-redundant database does not appear to be a major contributor to performance
223 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

246 small effective population size relative to other mammals (Leffler et al. 2012) and consequently a high
247 ratio of nonsynonymous to synonymous diversity (Fay et al. 2001; Kosiol et al. 2008). Thus,
248 distinguishing neutral and deleterious variants may be more difficult in humans than other species,
249 and approaches trained using human polymorphism may be more conservative with respect to
250 weakly deleterious variants. In comparison, predicting deleterious variants in *A. thaliana* may be
251 facilitated by the fact that it is a selfing species with an effective population size larger than that of
252 humans (Cao et al. 2011).

253 It should be noted that both demographic history and the process of local adaptation could
254 play important roles in the distribution of deleterious. In populations that are colonizing or expanding
255 into novel environments, the selective coefficients against individual variants may change (Slotte et al.
256 2013), and locally adaptive variants may become appreciably enriched. Both humans and *A. thaliana*
257 are known to have undergone demographic expansion in their recent evolutionary histories
258 (Hoffmann 2002; Finlayson 2005). While the relative extent of local adaptation in these two species is
259 difficult to quantify, both exhibit an excess of low frequency amino acid polymorphism characteristic
260 of deleterious variants (Lohmueller et al. 2008; Henn et al. 2016; Cao et al. 2011).

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

268 included genes with known mutant phenotypes, the sample of recently duplicated genes is limited.
269 Recent duplicates are more likely to accumulate common variants that would appear deleterious and
270 so may be among those genes where predictions are the most difficult.
271

272 Conclusions and future directions

273 Most approaches developed to predict deleterious mutations were trained using human data and in
274 many cases can only be used for human proteins, e.g., Kircher et al. 2014; Li et al. 2009; Schwarz et al.
275 2010. This study demonstrates that several generalized approaches perform exceptionally well in *A.*
276 *thaliana*, implying that they should also work well for other plant species. Despite the high
277 performance, it is quite likely further improvements could be achieved. Notably, LRT requires longer
278 run times than any of the other approaches, typically 5.2 hrs of compute time per gene. Although we
279 did not investigate whether a faster approach could be implemented without a loss in performance, it
280 is acknowledged that the long run time of the LRT may limit the application of the approach to large
281 genomics datasets. One potential avenue to pursue is whether faster heuristic measures of site-specific
282 conservation based on the BAD_Mutations pipeline of alignments could achieve similarly high
283 performance. However, further study would be needed to test whether heuristic measures of amino
284 acid conservation would be robust to the reference species and protein alignments to which they were
285 applied. A second approach would be to find a more effective means of generating predictions from
286 the combined output of multiple prediction approaches, as this has been shown to be highly effective
287 in humans, e.g. (González-Pérez and López-Bigas 2011). Although we did not find an ensemble
288 predictor that greatly improved performance, this might reflect the relatively small number of
289 variables used to generate ensemble predictions.

290

291 # Methods

292

293 Mutations with phenotypic effects were obtained from two sources. We generated a manually curated
294 set of 542 amino acid altering mutations in 155 genes with phenotypic effects that are described in the
295 literature. These mutations were found by searching the *Arabidopsis* Information Resource
296 (http://www.arabidopsis.org) for genes with either dominant or recessive alleles caused by nucleotide
297 substitutions. We also identified mutations using a literature search in Google Scholar
298 (http://scholar.google.com). For each variant we recorded the amino acid substitution, position and
299 link to the published paper (Table S3). We excluded nonsense mutations because they frequently
300 completely eliminate gene function. We identified a second set of 2,617 amino acid altering mutations
301 in 960 genes from the manually curated UniProt/Swiss-Prot database (http://www.uniprot.org/,
302 (Boutet et al. 2016). The two sets were independently generated and had an overlap of 249 mutants.
303 Using those mutants with named alleles as an indicator of those with gross versus biochemical
304 phenotypes, 65% of our manually curated set and 33% of the Swiss-Prot set had macroscopic
305 phenotypes. Duplicated genes were defined by those proteins with a significant blastp hit (E-value <
306 0.05) to another *A. thaliana* protein with greater than 60% identity. By this criteria 466/995 proteins
307 were classified as duplicated.

308 Single nucleotide polymorphisms (SNPs) without any known phenotype were obtained from a
309 set of 80 sequenced *A. thaliana* strains (Ensembl, version 81, "Cao_SNPs", (Cao et al. 2011)). At the
310 time of download, these were the only SNP set available with unrestricted use. After filtering out sites
311 with heterozygous or missing genotype calls, there were 10,797 biallelic amino acid altering SNPs in
312 the 995 proteins. We used a subset of 1,583 common SNPs (>10%) as those least likely to have

313 phenotypic effects.

314 We assessed amino acid substitutions using six approaches: LRT (Chun and Fay 2009),
315 PolyPhen2 (Adzhubei et al. 2010) , SIFT 4G (Vaser et al. 2016), Provean (Choi et al. 2012), MAPP(Stone
316 and Sidow 2005) and Gerp++ (Davydov et al. 2010). PolyPhen2 predictions were generated using the
317 standalone software (v2.2.2) with the PolyPhen2 bundled non-redundant database (uniref100-release
318 2011_12) and the probabilistic variant classifier using the default HumDiv model. Precomputed SIFT
319 4G predictions were obtained for *A. thaliana* (TAIR10.23) (http://sift.bii.a-star.edu.sg) and are based on
320 the UniRef90 database (2011). SIFT 4G predictions were not available for 855 substitutions,
321 predominantly because the amino acid change involved more than one nucleotide change within a
322 codon. Provean predictions (v1.1.5) were generated for all mutations using NCBI's non-redundant
323 database (04/02/2016). MAPP predictions were generated using BAD_Mutations alignments and trees
324 (see below). GERP++ generates predictions for single nucleotide positions rather than codons. To
325 assess GERP++ performance we used the GERP++ score at the first, second or third position of the
326 codon if the amino acid substitution could occur by a single change at one of those positions and the
327 average of the GERP++ scores at the first and second positions for all other types of changes. In
328 addition, because GERP++ did not perform well using neutral substitution rates estimated from each
329 alignment (default) we used a uniform neutral rate of 10 substitutions per site across all genes.

330 Predictions using a likelihood ratio test (LRT) were performed with the BAD_Mutations
331 pipeline (Kono et al. 2016). The pipeline makes use of sequenced and annotated genomes. We used
332 blast searches of 42 angiosperm genomes and retaining the top hit from each with a blast e-value
333 threshold of 0.05. Only Angiosperms were used to avoid extensive saturation of synonymous sites.
334 Pasta protein alignments (Mirarab et al. 2015) were generated using the homologs and the likelihood

335 of $dN = \omega dS$ compared to $dN = dS$ for each codon of interest was calculated using HYPHY (Pond et al.
336 2005), where dN and dS are the nonsynonymous and synonymous substitution rate and ω is a free
337 parameter. Sequences with 'N's or other ambiguous nucleotides were discarded prior to the likelihood
338 ratio test. The LRT differs compared to its original formation (Chun and Fay 2009) in that: i) dS was
339 estimated using all codons for each gene separately, ii) query sequences were optionally masked in
340 the likelihood calculation to avoid any reference bias and iii) branches with dS greater than 3 were set
341 to 3 to avoid spuriously high estimates of dS. Additionally, the original LRT used heuristics to
342 eliminate sites with $dN > dS$, the derived allele present in another species, or with fewer than 10
343 species in the alignment. Rather than eliminating sites, we used logistic regression to provide a single
344 probability of being deleterious based on the LRT test and these additional pieces of information.
345 Logistic regression was applied using both the masked and unmasked LRT p-values, where
346 the masked p-values were generated from alignments without the *A. thaliana* reference allele. For the
347 unmasked logistic regression, we used the terms log10(LRT p-value), constraint (dN/dS), Rn and An,
348 where Rn and An are the number of *A. thaliana* reference and alternative (i.e., mutant) amino acids
349 observed in the alignment, respectively. For the masked model we replaced An and Rn with the
350 absolute value of Rn – An and the maximum of Rn and An, respectively. For both models p-values
351 less than 1e-16 were set to 1e-16 and constraint values greater than 10 were set to 10. Ten-fold cross
352 validation was used to assess the fit of the logistic regression. The average area under the ROC curve
353 based on cross validation was 0.9575 (unmasked) and 0.9471 (masked). Because these values were
354 nearly identical to the performance of the model fit to the entire dataset, 0.9581 (unmasked) and 0.9471
355 (masked), we used the logistic regression coefficients from the full dataset:
356

357 log(p/(1-p)) = -2.407-0.2139*LRT(unmasked)-0.2056*constraint+0.07368*Rn-0.1236*An

358 log(p/(1-p)) = -2.453-0.1904*LRT(masked)-0.1459*constraint+0.2199*max(Rn,An)-0.2951*abs(Rn-An)

359

360 Sensitivity, specificity and area under the curve (AUC) were calculated for each approach

361 using the pROC package in R (Robin et al. 2011). Confidence intervals for each were calculated by

362 stratified bootstrapping (n = 2000).

363 Combined predictions were generated based on the combined scores of six approaches: LRT,

364 LRT-masked, PolyPhen2, Provean, GERP++ and MAPP. Sites with missing predictions from one or

365 more approach (n = 215) were removed. Combined predictions were generated using: 1) logistic

366 regression with each approach's score as a predictive variable, 2) support vector machine, 3) random

367 forest, 4) linear discriminant analysis and 5) generalized linear model with penalized maximum

368 likelihood implemented by the glmnet package in R (Friedman et al. 2010). The performance of each

369 model was assessed by AUC values obtained from 10-fold cross-validation.

370

371 # Data access

372 LRT predictions were implemented in the Python package BAD_Mutations which is freely available

373 from http://github.com/MorrellLAB/BAD_Mutations.git.

374

375 # Acknowledgements

376

377 We acknowledge funding support from the US National Science Foundation Plant Genome Program

378 grant (DBI-1339393 to JCF and PLM), from a University of Minnesota Doctoral Dissertation

379 Fellowship supporting TJYK. We thank members of the Morrell Lab for discussion and software

380 testing. Finally, we would like to thank Dr. Danelle Seymour from UC Irvine for comments.

381 **Author contributions:**

382 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.,
383 J.C.F., T.J.Y.K., and P.L.M. wrote the initial draft of the manuscript. All authors contributed to final
384 manuscript preparation.

385

386 # Reference:

387 1000 Genomes Project Consortium T 1000 GP. 2012. An integrated map of genetic variation from 1,092
388 human genomes. *Nature* **491**: 56–65.

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

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

393 Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hébert S, Doelle H, Ersoy B, Kryukov
394 G, Schmidt S, et al. 2007. Medical sequencing at the extremes of human body mass. *Am J Hum Genet*
395 **80**: 779–791.

396 Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios
397 I. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to
398 use the entry view. In *Plant Bioinformatics* (ed. D. Edwards), Vol. 1374 of, pp. 23–54, Springer New
399 York, New York, NY http://link.springer.com/10.1007/978-1-4939-3167-5_2 (Accessed January 19,
400 2017).

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

403 Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C,
404 et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**: 956–
405 963.

406 Charlesworth B. 2012. The role of background selection in shaping patterns of molecular evolution
407 and variation: evidence from variability on the *Drosophila* X chromosome. *Genetics* **191**: 233–246.

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

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

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

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

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

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

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

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

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

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

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

430 Fay JC, Wyckoff GJ, Wu C-I. 2001. Positive and negative selection on the human genome. *Genetics* **158**:
431 1227–1234.

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

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

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

437 Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U,

438 Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*
439 **40**: D1178–D1186.

440 Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson
441 PD, Daly MJ, Smoller JW, et al. 2015. The evaluation of tools used to predict the impact of missense
442 variants is hindered by two types of circularity. *Hum Mutat* **36**: 513–523.

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

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

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

450 Hoffmann MH. 2002. Biogeography of *Arabidopsis thaliana* L. Heynh. (Brassicaceae). *J Biogeogr* **29**: 125–
451 134.

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

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

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

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

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

463 Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M.
464 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLOS Biol* **10**:
465 e1001388.

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

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

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

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

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

478 Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:
479 1151–1155.

480 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM,
481 Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature*
482 **461**: 747–753.

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

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

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

489 Morrell PL, Buckler ES, Ross-Ibarra J. 2012. Crop genomics: advances and applications. *Nat Rev Genet*
490 **13**: 85–96.

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

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

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

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

499 Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**:
500 676–679.

501 Renaut S, Rieseberg LH. 2015. The accumulation of deleterious mutations as a consequence of
502 domestication and improvement in sunflowers and other compositae crops. *Mol Biol Evol* **32**: 2273–
503 2283.

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

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

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

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

513 Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS,

514 Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating
515 system evolution. *Nat Genet* **45**: 831–835.

516 Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. 2014. The Human Gene Mutation
517 Database: building a comprehensive mutation repository for clinical and molecular genetics,
518 diagnostic testing and personalized genomic medicine. *Hum Genet* **133**: 1–9.

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

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

523 Thornton KR, Foran AJ, Long AD. 2013. Properties and modeling of GWAS when complex disease
524 risk is due to non-complementing, deleterious mutations in genes of large effect. *PLOS Genet* **9**:
525 e1003258.

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

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

530 Yang J, Mezmouk S, Baumgarten A, Buckler ES, Guill KE, McMullen MD, Mumm RH, Ross-Ibarra J.
531 2016. Incomplete dominance of deleterious alleles contribute substantially to trait variation and
532 heterosis in maize. *bioRxiv* 86132.

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

535

536 ## Tables

537

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

538 * Highest performing approach for a given test set.

539 [1] LRT in this study used a different alignment pipeline than the LRT applied to the human test sets.

540

541

542 ## Figure Legends

543

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

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

548

549 Figure 2. The proportion of SNPs called deleterious across frequency classes. The fraction of SNPs

550 called deleterious by each approach (legend) at its 95% specificity threshold across five frequency

551 classes, labeled by the number of minor alleles present (*n* =80). Sample sizes for the five classes are

552 5303 (1), 1646 (2), 1250 (3-4), 1015 (5-8) and 1583 (>8).

553

554 Figure 3. Performance of approaches across different classes of sites. Performance is measured by the

555 area under the curve (AUC) of the approach's sensitivity versus specificity. A – comparison of

556 mutants with biochemical versus gross phenotypes. B – comparison of performance for duplicated

557 versus single copy genes.

558

559 Figure 4. Dissimilarities among approaches. Dissimilarities were computed by the pairwise number of

560 disagreements between each approach applied to mutants and common SNPs. Dissimilarities are

561 represented by a tree based on hierarchical clustering.

562

563 Supplemental Tables

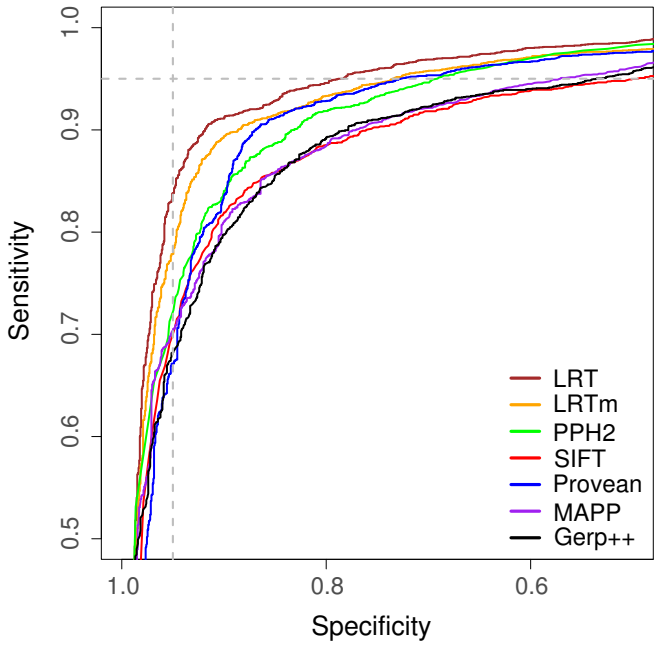
564

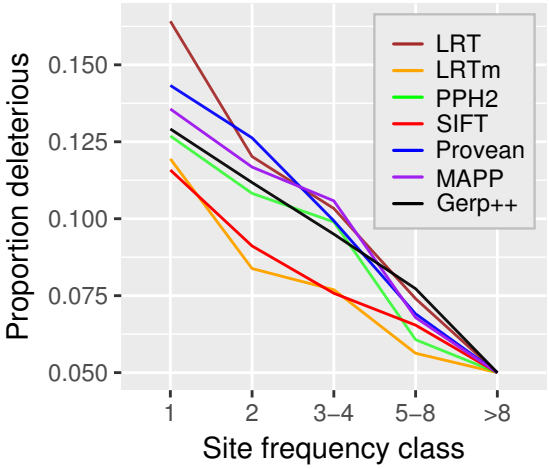
565 Table S1. Performance of methods used to distinguish deleterious and neutral substitutions

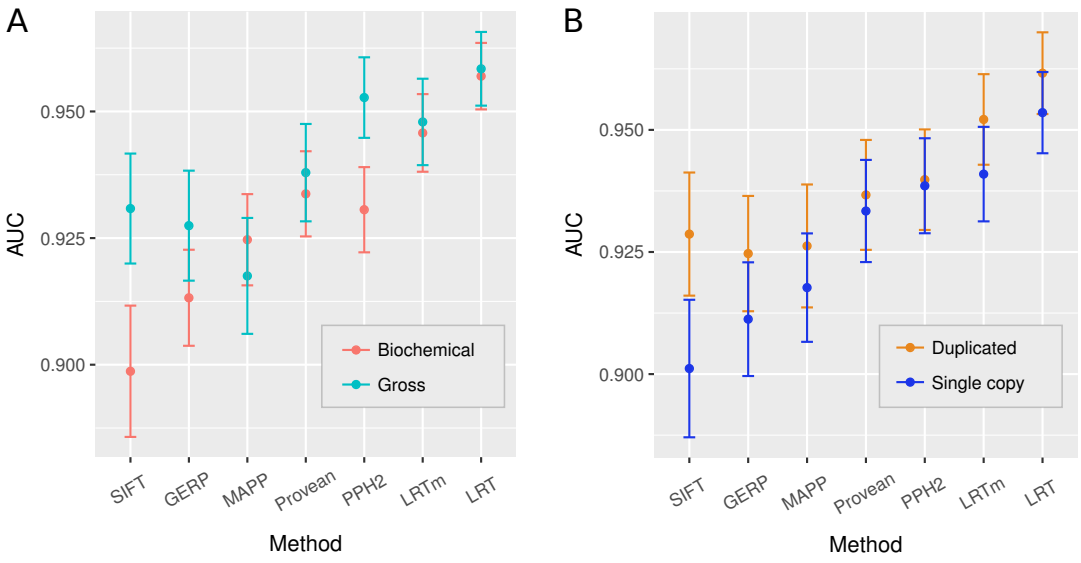
566 Table S2. Performance of models based on ensemble prediction methods.

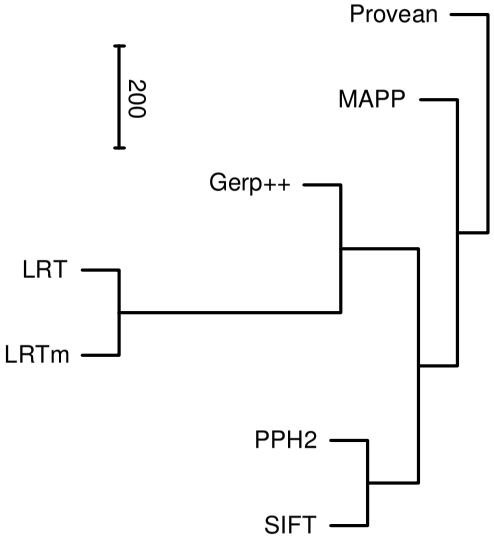
567 Table S3. 2,617 amino acid altering mutations in 960 *A. thaliana* genes

568


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


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


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


LRT
LRTm
200
Gerp++
PPH2
SIFT
MAPP
Provean