Theor Appl Genet (2010) 121:157–168
DOI 10.1007/s00122-010-1299-4

ORIGINAL PAPER

Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice

Torsten Günther · Karl J. Schmid

Received: 11 May 2009 / Accepted: 11 February 2010 / Published online: 3 March 2010

Abstract Plant genetic diversity has been mainly investigated with neutral markers, but large-scale DNA sequencing projects now enable the identification and analysis of different classes of genetic polymorphisms, such as non-synonymous single nucleotide polymorphisms (nsSNPs) in protein coding sequences. Using the SIFT and MAPP programs to predict whether nsSNPs are tolerated (i.e., effectively neutral) or deleterious for protein function, genome-wide nsSNP data from *Arabidopsis thaliana* and rice were analyzed. In both species, about 20% of polymorphic sites with nsSNPs were classified as deleterious; they segregate at lower allele frequencies than tolerated nsSNPs due to purifying selection. Furthermore, *A. thaliana* accessions from marginal populations show a higher relative proportion of deleterious nsSNPs, which likely reflects differential selection or demographic effects in subpopulations. To evaluate the sensitivity of predictions, genes from model and crop plants with known functional effects of nsSNPs were inferred with the algorithms. The programs predicted about 70% of nsSNPs correctly as tolerated or deleterious, i.e., as having a functional effect. Forward-in-time simulations of bottleneck and domestication models indicated a high power to detect demographic effects on nsSNP frequencies in sufficiently large datasets. The results indicate that nsSNPs are useful markers for analyzing genetic diversity in plant genetic resources and breeding populations to infer natural/artificial selection and genetic drift.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1299-4) contains supplementary material, which is available to authorized users.

Communicated by A. Schulman.

T. Günther · K. J. Schmid
Leibniz Institute of Plant Genetics
and Crop Plant Research (IPK), Gatersleben, Germany

Present Address:
T. Günther · K. J. Schmid (✉)
Institute of Plant Breeding, Seed Science
and Population Genetics, University of Hohenheim,
Fruwirthstrasse 21, 70599 Stuttgart, Germany
e-mail: karl.schmid@uni-hohenheim.de

K. J. Schmid
Swedish University of Agricultural Sciences (SLU),
Uppsala, Sweden

Introduction

The analysis of genetic variation in natural plant populations, the management of genetic resources, and the selection of individuals in breeding programs is frequently based on the analysis of supposedly neutral genetic markers, such as silent single nucleotide polymorphisms or simple sequence repeats. Meanwhile, the large-scale analysis of plant genomic variation revealed a bewildering diversity of genetic polymorphisms. Point mutations, repeat variants, polyploidization, transposable element insertions, and gene duplications are abundant in plant genomes and contribute to phenotypic variation in wild and crop plants (e.g., Fu et al. 2002; Morgante et al. 2005). This diversity of genetic variants raises the question of whether they differ in their effect on plant fitness and phenotypic diversity, and how they can be utilized for genetic mapping and plant breeding.

The evolutionary fate of a genetic polymorphism depends on its fitness effects, which is expressed by the selection coefficient, s, and the effective population size, N_e (Kimura and Crow 1963). Strongly deleterious polymorphisms are rapidly removed by selection, but slightly

deleterious polymorphisms with $s \approx 1/N_e$ accumulate by genetic drift and reduce average population fitness (Kimura 1962; Lande 1994). In wild plant species, environmental changes, ecological specialization, or limited dispersal cause small effective population sizes. Artificial selection and inbreeding during domestication reduced levels of genetic variation in crop plants. The combination of artificial selection, hitchhiking and low variation from high levels of inbreeding during domestication is known as the domestication-associated Hill-Robertson effect (Lu et al. 2006; Yamasaki et al. 2007). For example, Wright et al. (2005) estimated that 2–4% of maize genes show reduced variation as a consequence of artificial selection. Modern breeding populations may accumulate slightly deleterious mutations by genetic drift, by linkage to advantageous yield improvement genes, or because irrigation and fertilization buffer their fitness effects. One consequence of this process is an increased susceptibility for diseases in elite varieties (Gepts and Papa 2002).

A significant portion of phenotypic variation in plants is caused by single nucleotide polymorphisms (SNPs) that alter gene regulation, gene function, or mRNA splicing patterns (e.g. Johanson et al. 2000; Maloof et al. 2001; Cartegni et al. 2002; Stein et al. 2005; Filiault et al. 2008). Within coding regions, SNPs are differentiated into synonymous, non-synonymous, and nonsense SNPs. While synonymous SNPs do not affect protein sequence, non-synonymous SNPs (nsSNPs) cause amino acid polymorphisms and nonsense mutations lead to a premature stop codon. In populations of both wild and crop plants (Nordborg et al 2005; Schmid et al. 2005; Hamblin et al. 2006; Lu et al. 2006), nsSNPs segregate at a lower frequency than synonymous SNPs, suggesting genome-wide purifying selection against deleterious amino acid polymorphisms. However, not all amino acid polymorphisms are deleterious because their effects on protein function depend on the location in the protein structure and the physico-chemical distance between polymorphic amino acids (Suckow et al. 1996). Several methods were developed to predict whether nsSNPs have deleterious effects on protein function (reviewed in Ng and Henikoff 2006). They analyze the evolutionary conservation of protein sequence or the location of amino acid polymorphism in the protein structure.

For this paper we quantified the relative frequency of deleterious nsSNPs in plants and tested whether nsSNPs are useful markers for the analysis of natural selection in wild and crop plants. We used publicly available genome-wide datasets from *Arabidopsis thaliana* and rice to characterize variation in the ratio of deleterious to tolerated nsSNPs on a genome-wide level. Based on the prediction results, we investigated the effect of demographic history on the efficiency of purifying selection against deleterious amino acid polymorphisms. To validate the results of prediction programs for individual genes, we compared predicted fitness effects of nsSNPs with functional studies of more than a dozen functionally well-characterized genes from various plant species. Finally, we conducted simulations to test the power of prediction programs to detect changes in deleterious nsSNP frequency after population bottlenecks.

Materials and methods

Genome-wide datasets of nsSNP polymorphism

Two publicly available datasets from *Arabidopsis thaliana* and one dataset from rice were analyzed. The first dataset ('2010' thereafter) comprised annotated DNA sequence polymorphism data from 778 protein coding regions in 96 natural *A. thaliana* accessions (Nordborg et al. 2005). The Van-0 accession was excluded because of a high level of heterozygosity (Nordborg et al. 2005). Homologous sequences from the closely related species *Arabidopsis lyrata* obtained from http://www.jgi.doe.gov/Alyrata were used as outgroup to distinguish between derived and ancestral alleles. Orthologous sequences in *A. lyrata* were identified by reciprocal BLAST searches. The second *A. thaliana* dataset ('Perlegen') included 121,418 nsSNPs identified by re-sequencing of 20 accessions (Clark et al. 2007). The rice dataset consisted of DNA sequence polymorphism data at 111 loci from 97 accessions of wild and domesticated rice populations (Caicedo et al. 2007). Sequences were obtained from cultivated rice *Oryza sativa*, its wild ancestor *Oryza rufipogon*, and other wild rice species, such as *Oryza nivara*, *Oryza barthii*, and *Oryza meridionalis*. To annotate rice sequences, a BLASTX comparison against protein sequences from TIGR rice genome annotation 5 was conducted (TIGR 2007) and best hits were used to predict the coding sequence with Wise 2.2.0 (Birney et al. 2004). *O. barthii* or *O. meridionalis* were used as outgroups.

To further test prediction methods, they were applied to nsSNPs with experimentally inferred functional effects in model and crop plants (see Table 2).

Prediction of functional effects of amino acid polymorphisms

Functional effects of amino acid polymorphisms on protein function (tolerated versus deleterious) were predicted with SIFT 2.1.2 (Ng and Henikoff 2001, 2003) and MAPP (Stone and Sidow 2005) programs. SIFT uses PSI-BLAST (Altschul et al. 1997) to identify homologous proteins in a database. We used the TrEMBL 37.4 database (Boeckmann

et al. 2003) for homolog identification. MAPP requires a collection of homologous proteins as well, and for that reason PSI-BLAST alignments from the TrEMBL database searches were used for both SIFT and MAPP analyses. Furthermore, MAPP requires a phylogenetic tree of homologous sequences as input. The trees were constructed from PSI-BLAST alignments of protein sequences with the neighbor joining method implemented in the SEMPHY 2.00 program (Friedman et al. 2002). Sequences were translated with BioPython 1.43 (http://www.biopython.org) and aligned with the reference sequence using PAL2NAL (Suyama et al. 2006) to obtain a codon-based nucleotide alignment.

Forward-in-time simulations of evolutionary scenarios

To investigate the effect of demographic history on the relative frequency of deleterious nsSNPs, forward-in-time simulations were conducted. Sequences derived from the *lacI* gene of *E. coli* were used as starting sequences for the simulation; the functional effects of amino acid substitutions in this protein are known from mutagenesis studies (Suckow et al. 1996). Based on inferred effects, a fitness value was assigned to individuals in the population by adding up positive and negative fitness values of nsSNPs. In each generation, sequences were mutated and individuals for the next generation were selected according to their fitness value. Details of the simulation procedure are described in the Supplementary Information.

We simulated a model with a population bottleneck using the following model parameters (Fig. 1). The starting population size, N_0, was set to 10,000 for 5,000 generations. Then, population size was reduced to 100 individuals for 250 generations. After the bottleneck, the population size, N_2 was increased to 10,000 individuals and the simulation was run for another 7,250 generations. These parameter values correspond to the domestication model of Innan and Kim (2004). The nucleotide substitution rate of plants was estimated to range from 5 to 30×10^{-9}, although the true rate is probably closer to the lower boundary (Wolfe et al. 1987). Therefore, a mutation rate $\mu = 10^{-8}$ was chosen. To ensure that $|s| \ll 1/N_e$ and the fixation of a polymorphism by drift can occur during the bottleneck, a fitness penalty (see Supplementary Information) was set to $P = 10^{-6}$, corresponding to $|s| \in [10^{-6}, 10^{-5}]$.

We compared a null model with a test model and explored two scenarios. The first scenario describes a simple bottleneck in one population and a constant size in the other; in both populations, all loci evolved under selection. The second scenario was identical to the first one with the exception that selection acted on only one of the 10 simulated genes and the others were allowed to evolve neutrally. In the null model of this scenario, all loci evolved under selection. Simulations of each scenario were repeated 1,000 times and conducted with and without recombination between loci.

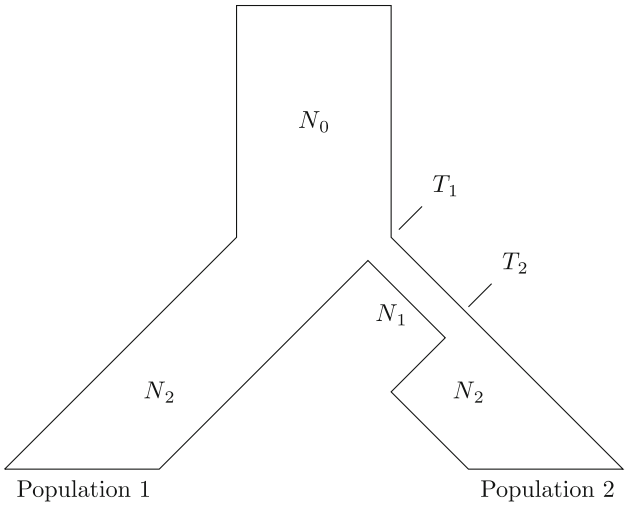

Fig. 1 Demographic model used for the simulation of the bottleneck and domestication models. Population size was reduced to N_1 between generations T_1 and T_2, and later increased to the previous size, N_0

Programs were written in Python and statistical tests were conducted with GNU R 2.5.1.

Results

Comparison of the SIFT and MAPP programs

We first compared prediction results of SIFT and MAPP programs, which are summarized in Table 1. In the three datasets combined, a total of 65,355 nsSNPs were analyzed. The SIFT program made predictions for 47,305 and MAPP for 47,487 nsSNPs, and predictions were identical for 87.3% of polymorphisms (Supplementary Information). Furthermore, there was a high correlation between SIFT and MAPP prediction scores (Spearman's $r_s = 0.72$). The correlation of either SIFT and MAPP predictions with an analysis using a BLOSUM62 matrix was much lower ($r_s = 0.28$ and 0.30, respectively). There were no significant differences in the distribution of both categories (deleterious vs. tolerated) between the three datasets within each program (G test, SIFT: $P = 0.080$, MAPP: $P = 0.583$). These comparisons indicate that SIFT and MAPP predictions are quite similar, although they differ in details as outlined below. In the following sections we present the MAPP predictions and show the SIFT results only if they differ from the MAPP output.

Table 1 Summary of predictions with SIFT and MAPP

Prediction method	*Arabidopsis thaliana*				Rice	
	2010		Perlegen			
	nsSNPs	%	nsSNPs	%	nsSNPs	%
SIFT						
Total	1,128	100.0	46,012	100.0	165	100
Tolerated	872	77.3	36,428	79.2	139	84.2
Deleterious	256	22.7	9,584	20.8	26	15.8
MAPP						
Total	919	100.0	46,402	100.0	166	100.0
Tolerated	756	82.3	37,571	81.0	133	80.1
Deleterious	163	17.7	8,831	19.0	33	19.9

Non-synonymous SNP frequencies in *Arabidopsis thaliana*

The genome-wide dataset from the 2010 dataset comprises 383 loci from *A. thaliana* that could be aligned to orthologous genes from *A. lyrata*. The data included 1,174 nsSNPs, of which 919 were predicted by MAPP. Figure 2 shows the frequency distribution of deleterious nsSNPs among *A. thaliana* accessions. The SIFT prediction for Col-0 shows a significantly lower proportion of deleterious nsSNPs compared to all other accessions (*G* test, $P < 0.001$, Fig. 2a). In the MAPP analysis, there was no difference between Col-0 and the other accessions (*G* test, $P = 0.4125$; Fig. 2b). The SIFT algorithm assumes that the external BLAST database contains solely tolerated nsSNPs. Since Col-0 was used as reference sequence for the *A. thaliana* genome project, its sequences are included in external databases and polymorphisms in Col-0 are classified as tolerated by SIFT. This suggests to remove reference sequences from the alignment input for SIFT. The MAPP algorithm does not make such an assumption and is unbiased with respect to the reference sequence.

The frequency distributions of nsSNPs in the complete 2010 dataset show that derived nsSNPs with a predicted deleterious effect segregate at a lower allele frequencies than tolerated nsSNPs (Fig. 3a; Kolmogorov–Smirnov Test, $P = 0.007$; Supplementary Information). The ratio of deleterious to tolerated nsSNPs decreases linearly with derived nsSNP frequencies (Fig. 3d).

The Perlegen data included 121,418 SNPs from 20 *A. thaliana* accessions, of which 63,949 SNPs were retained for further analysis after alignment with *A. lyrata* orthologs. Predictions were obtained for 46,402 nsSNPs (Supplementary Information). In contrast to the 2010 data, the Col-0 accession exhibits a significantly lower proportion of deleterious nsSNPs than other accessions with both prediction programs (*G* test, $P < 0.0001$; Fig. 2c, d). The average proportion of deleterious nsSNP positions is 0.087 and 0.133 for the 2010 and Perlegen data, respectively, if the Col-0 accession is excluded. This difference is significant (*t* test, $P < 0.0001$). As in the 2010 data, derived

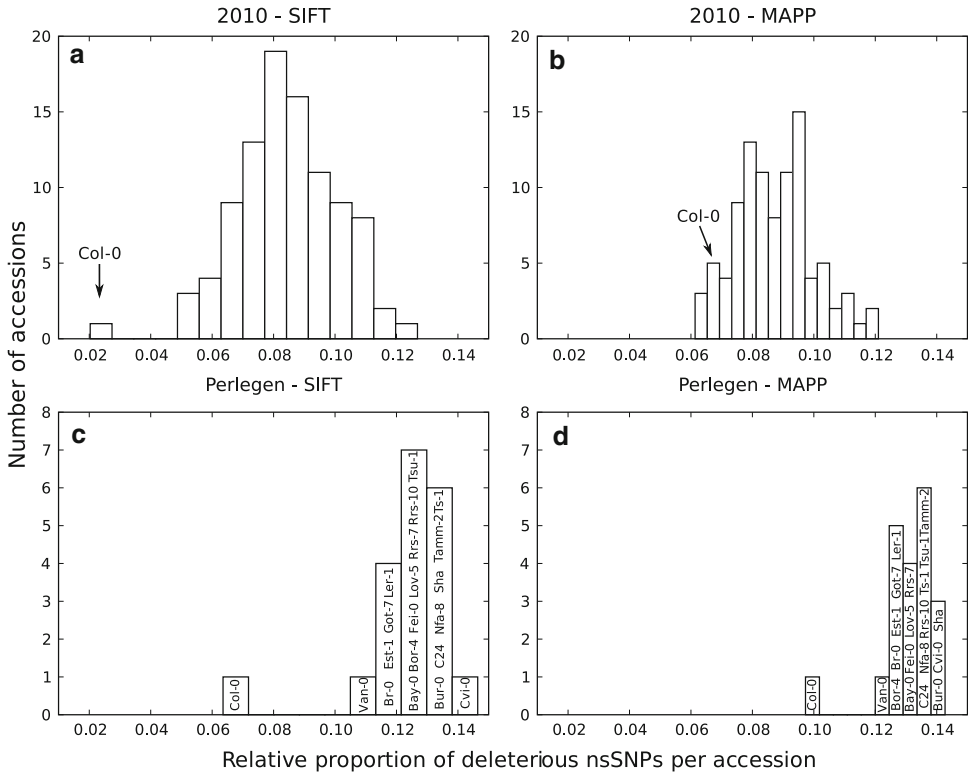

Fig. 2 Distribution of the relative proportions of deleterious nsSNPs among accessions in the 2010 (**a** and **b**) and Perlegen (**c** and **d**) datasets

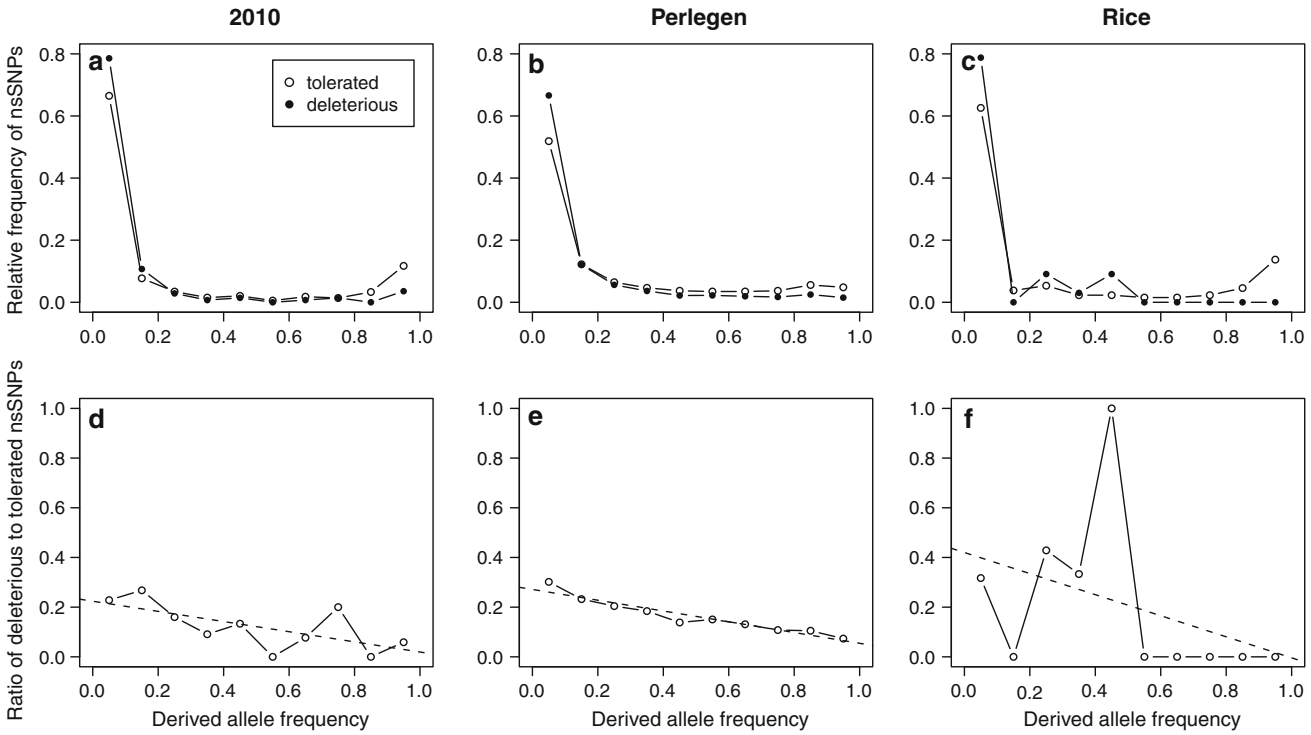

Fig. 3 Allele frequency distribution of tolerated and deleterious nsSNPs for the 2010 (**a**), Perlegen (**b**), and rice (**c**) datasets. The proportion of nsSNPs for each frequency class was calculated relative to the total number of nsSNPs in the corresponding fitness category. The ratio of absolute numbers of deleterious to tolerated nsSNPs of each dataset is shown in **d**, **e**, and **f**, respectively. The *dashed lines* represent the linear regression line

deleterious nsSNPs segregate at a lower allele frequency than tolerated nsSNPs (Fig. 3b; Kolmogorov–Smirnov test, $P < 0.0001$) and the ratio of deleterious to tolerated nsSNPs decreases linearly with derived SNP frequencies (Fig. 3e). The regression slopes of this decay are nearly identical in the 2010 (−0.205) and the Perlegen data (−0.217).

Population genetics theory predicts that purifying selection is less effective in smaller and endemic populations. To test whether the proportion of deleterious nsSNPs differs between populations, we grouped the accessions into monophyletic clades according to Supplementary Fig. 1 of Nordborg et al. (2005). The Van-0 and Col-0 accessions were not included for reasons discussed above. The distribution of deleterious and tolerated nsSNPs was compared in the remaining clades (Supplementary Table 5). We observed that genetically divergent accessions, such as Cvi-0, Shahdara, C24, and Bur-0 contain a significantly higher proportion of deleterious amino acid polymorphisms than other accessions. The results are consistent between the 2010 and Perlegen data but the latter data showed more significant results because of a larger number of nsSNPs.

Since the Perlegen data covers the complete genome, we tested whether gene families differ by their proportions of deleterious nsSNPs. As has been noted before (Clark et al. 2007), the average number of nsSNPs per gene is highly variable between gene families (Fig. 4). The NBS-LRR and receptor-like kinases are the most polymorphic, and the cytoplasmatic ribosomal proteins the least polymorphic gene families. In contrast, the proportion of putatively deleterious polymorphisms does not vary much between gene families. Only the conserved gene family of ribosomal cytoplasmatic proteins contains a significantly increased proportion of nsSNPs (G test, $P = 0.005$); bHLH transcription factors exhibit a significantly decreased proportion of deleterious nsSNPs ($P = 0.03$).

Non-synonymous SNP frequencies in rice

Domesticated species may accumulate slightly deleterious mutations because of relaxed purifying selection. To test this hypothesis, we analyzed a dataset from different rice species with 232 nsSNPs and 166 MAPP predictions (Table 1 and Supplementary Information). As in the *A. thaliana* data, there was an excess of rare over high-frequency nsSNPs, but allele frequency distributions of deleterious and tolerated nsSNPs in rice were not significantly different (Kolmogorov–Smirnov test, $P = 0.1044$; Fig. 3c). In comparison to wild *O. rufipogon*, both domesticated rice species *O. sativa* ssp. *japonica* and *O. sativa* ssp. *indica* show a higher proportion

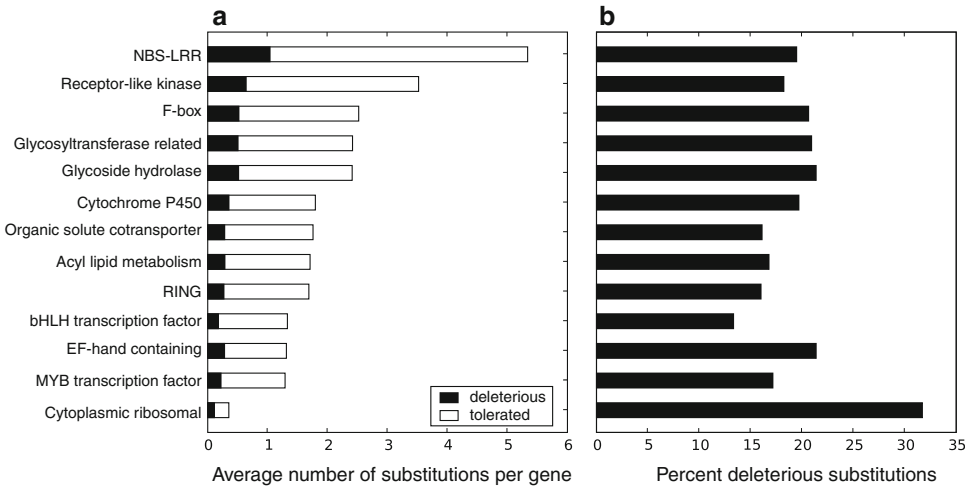

Fig. 4 Differences in the relative proportion of deleterious nsSNPs among protein families based on MAPP predictions

of deleterious nsSNPs. However, these differences are not statistically significant due to the small sample sizes (G test, $P = 0.7$ and $P = 0.4$). Figure S1 from Caicedo et al. (2007) shows a bifurcation between the monophyletic groups of wild rice *O. rufipogon* from different origins and the two domesticated species *O. sativa* ssp. *japonica* and *O. sativa* ssp. *indica* together with all *O. rufipogon* accessions from China. Such a pattern results from two independent domestication events in China (Kovach et al. 2007). Testing this phylogenetic grouping revealed a significantly higher proportion of deleterious nsSNPs in the group containing the domesticated rice species compared to the non-Chinese *O. rufipogon* accessions (G test, $P = 0.045$). Instead of a linear decay in the relative frequency of deleterious polymorphisms we observed a peak at intermediate frequency nsSNPs (Fig. 3f). This pattern may be interpreted as a domestication-associated enrichment of deleterious mutations. However, since the total number of nsSNPs in this analysis is quite low, it could represent random noise.

Prediction results for polymorphisms with an inferred phenotype

To test the sensitivity of the predictions with the SIFT and MAPP programs, we assembled a dataset from the literature that comprised 68 nsSNPs with experimentally inferred functional effects. Predictions of functional effects were obtained for 53 nsSNPs (Table 2). As much as 8 out of the 15 substitutions without a prediction occur in the *A. thaliana PHYB* flowering control gene (Filiault et al. 2008) and all of them are associated with phenotypic changes. They are located in highly variable regions of the protein and no prediction was possible because of a low alignment quality of these regions. The same problem was also observed in other proteins and demonstrates that sequence-based prediction methods are restricted to protein regions that can be aligned reliably.

Among 53 nsSNPs with a MAPP prediction, 20 were found to have a significant effect and 14 to have no effect on protein function (Table 2). The functional effects of the remaining 19 substitutions are unknown because their function was not studied or no functional effect was found. In some cases, groups of nsSNPs were shown to have an effect on protein function like in the barley *eIF4E* gene (Stein et al. 2005). In the complete dataset, MAPP classified a single neutral nsSNP as deleterious, which corresponds to a false positive rate (FPR) of 7.1% (1 out of 14). The false negative rate (FNR) was higher because only 11 out of 20 nsSNPs (45% FNR) with functional effects were predicted correctly.

Forward-in-time simulations of demographic models

The neutral theory of molecular evolution states that the fixation probability of slightly deleterious polymorphisms depends on the effective population size (Kimura 1962). Therefore, we hypothesized that the observed ratio of deleterious to tolerated nsSNPs provides information on the extent of purifying selection acting on a population. We used forward simulations to test the power of detecting the effect of bottlenecks on the frequency of deleterious nsSNPs. The fitness assessment was based on empirical data from mutagenesis experiments of the *lacI* gene in *E. coli*. Even though fitness effects of mutations likely differ between genes (Li 1997), we assumed that *LacI* can be taken as a representative protein.

On average, three nsSNPs were observed per locus per individual in the simulated populations. Power was measured by counting the number of significant G tests of homogeneity ($P < 0.01$) for differences in the proportion of deleterious to tolerated nsSNPs between bottlenecked and non-bottlenecked populations. Tests were carried out with different sample sizes (10, 100, 1000, or 10,000 individuals from both populations), but with a constant

Table 2 Prediction results for nsSNPs with inferred phenotypic effects

Gene	nsSNP	SIFT	MAPP	Phenotype
CRY2	Flowering time in *Arabidopsis thaliana*; GenBank ID: AAD09837.1 (El-Assal et al. 2001)			
	Q127S	+	+	No effect
	S188L	+	+	No effect
	V367M	−	−	Changed flowering time
	V476I	+	+	No effect
eIF4E	Bymovirus resistance in *Hordeum vulgare*; CAG27836.1 (Stein et al. 2005)			
	S57F	−	+	*rym4*-specific
	K118T	+	−	*rym4*-specific in European samples
	K118I	−	−	*rym4*-specific in Asian samples
	T120S	+	+	*rym5*-mediated resistance
	N160D	−	−	*rym5*-mediated resistance
	Q161K	+	+	*rym4*-specific
	S205F	+	+	*rym4*-specific in European samples
	D206G	+	+	*rym4*-specific
	G208A	+	+	No effect
	G208S	+	+	No effect
eIF4E	Potyvirus resistance in *Capsicum*; AAR23916.1 (Yeam et al. 2007)			
	T51A	+	+	No effect
	V67E	+	−	Reduced susceptibility
	L79R	−	−	Reduced susceptibility
	G107R	+	−	Resistance
	D109N	−	−	No effect
eIF4G	Rice yellow mottle virus resistance in *O. sativa*; CAJ42897.1 (Albar et al. 2006)			
	E309K	−	+	Resistance
FRI	Flowering time in *A. thaliana*; AAG23415.1 (Gazzani et al. 2003; Johanson et al. 2000)			
	F55I	+	+	No effect
	R74C	−	+	Potentially loss of function
	D167E	+	+	Potentially loss of function
	K377Q	+	+	Behind nonsense-mutation
	L79I	+	+	No effect
	G146E	+	−	Earlier flowering time (linked to M148I)
	M148I	+	+	Earlier flowering time (linked to G146E)
GI	Fruit set in *A. thaliana*; NP_564180.1 (Brock et al. 2007)			
	G672R	+	−	Reduced fruit set
Isa	Climatic adaptation in *Hordeum spontaneum*; CAA78305.1 (Cronin et al. 2007)			
	R22C	+	+	Correlated to latitude and seasonal temperature difference
	A51S	+	+	
	H65L	+	+	Correlated to diurnal temperature difference and evaporation
	V90A	+	+	
	S93P	−	−	
	I96V	+	+	
OsC1	Anthocyanin pigmentation in *O. sativa*; BAD04024.1 (Saitoh et al. 2004)			
	P41R	−	−	Reduced pigmentation
	S90F	−	−	Reduced pigmentation
	P148Q	+	+	No effect
PHYA	Light sensitivity in *A. thaliana*; NP_172428.1 (Maloof et al. 2001)			
	M548T	−	−	Reduced far red light sensitivity

Table 2 continued

Gene	nsSNP	SIFT	MAPP	Phenotype
PHYB	Reaction to light in *A. thaliana*; NP_179469.1 (Filiault et al. 2008)			
	A247S	+	+	No significant correlation to light response
	A624D	+	+	
	E709K	+	+	
	S736T	−	+	No significant correlation to light response
	K742T	−	−	
RCY1	Cucumber mosaic virus resistance in *A. thaliana*; BAC67706.1 (Sekine et al. 2006)			
	D47N	−	−	Loss of function in resistance gene
	W217C	−	−	Loss of function in resistance gene
	R550K	−	+	Loss of function in resistance gene
	E673K	−	−	Loss of function in resistance gene
RTM1	Tobacco etch virus resistance in *A. thaliana*; NP_172067.1 (Chisholm et al. 2000)			
	S56F	−	−	Loss of resistance
	G132D	+	−	Loss of resistance
	D139N	+	+	Loss of resistance
sh4	Grain shattering in *O. sativa*; ABD35853.1 (Li et al. 2006)			
	K79N	+	+	Reduced grain shattering
WAP2	Free threshing in *Triticum aestivum*; AAU94926.1 (Simons et al. 2006)			
	V329I	+	+	Confers free threshing character

+ tolerated, − deleterious nsSNP

number of 10 genes. The power to detect differences between bottlenecked and non-bottlenecked populations increased with sample size and was generally higher in simulations without recombination (Fig. 5), due to increased genetic drift from stronger background selection in non-recombining populations (Charlesworth et al. 1993). The simulations revealed no significant power differences between the bottleneck and domestication (bottleneck and artificial selection) scenarios (Fig. 5).

Discussion

Signature of purifying selection in genome-wide data

All three empirical genome-wide datasets showed variable proportions of deleterious nsSNPs among accessions and allele frequencies. Overall, deleterious nsSNPs segregate at a lower frequency than tolerated nsSNPs. The class of rare nsSNPs with a frequency of <10% harbors a higher proportion of deleterious polymorphisms than high-frequency nsSNPs (Fig. 3). This pattern is consistent with either purifying or balancing selection. Under purifying selection, deleterious nsSNPs are selected against and remain at low frequency (Wong et al. 2003). Balancing selection may produce low-frequency polymorphisms if multiple alleles or haplotypes are favored by selection. This seems unlikely because *A. thaliana* and *O. sativa* are mainly self-fertilizing species with low levels of heterozygosity (Oka 1988; Abbott and Gomes 1989). Hence, balancing selection caused by heterozygote advantage is probably rare. Alternatively, local adaptation in self-fertilizing species can lead to patterns of genetic variation resembling balancing selection (Hedrick 1998). Some nsSNPs classified as deleterious may in fact cause advantageous functional changes in encoded proteins and evolve by positive or balancing selection. In this case, a higher proportion of 'deleterious' nsSNPs may be expected in gene families, such as disease resistance genes that likely are more frequently targets of positive selection than other gene families (Clark et al. 2007). However, a comparison of total nsSNP counts with the proportion of deleterious nsSNPs in different gene families shows that the proportion of deleterious polymorphisms is remarkably similar between protein families (Fig. 4).

We observed a nearly identical decrease in the relative frequency of deleterious nsSNPs with increasing nsSNP frequency in the two *A. thaliana* datasets (Fig. 3d, e). This pattern is noteworthy because the two datasets differ strongly in the numbers of SNPs and accessions. Since high-frequency SNPs are on average older than low-frequency SNPs, deleterious nsSNPs do not reach higher frequencies as often as tolerated amino acid polymorphisms. A plausible explanation is that deleterious polymorphisms are selected against by purifying selection. In this case, the slope of the decrease is influenced by the

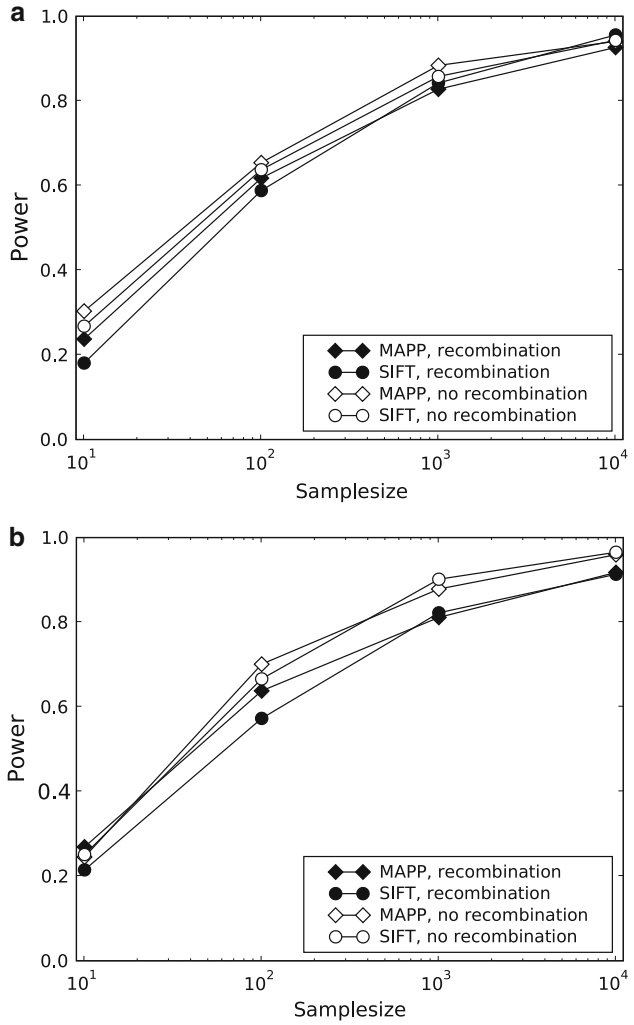

Fig. 5 Power to detect the effect of a bottleneck on the frequency of deleterious nsSNPs with SIFT and MAPP under a bottleneck model (**a**) and a domestication model (**b**), simulations were run with and without recombination between loci

strength of purifying selection and may allow to estimate the extent of purifying selection in populations and species of different effective population sizes. Taken together, our analyses support genome-wide acting purifying selection rather than gene-specific positive selection as an explanation for the lower frequency of deleterious nsSNPs.

Different proportions of deleterious nsSNPs in *A. thaliana* accessions

In *A. thaliana*, significant differences in relative frequencies of deleterious nsSNPs between groups of accessions were observed in the Perlegen dataset (Supplementary Table 5). In particular, the Cvi-0 accession exhibits an increased proportion of deleterious nsSNPs compared to other accessions. Cvi-0 originated from Cape Verde, a small group of islands that is located 500 km away from the mainland at the edge of the species distribution (Hoffmann 2002). The high frequency of deleterious nsSNPs is consistent with its phenotypic and genetic divergence from other *A. thaliana* accessions (Alonso-Blanco et al. 1999; Schmid et al. 2003; Nordborg et al. 2005). The Shahdara accession, which originated from an isolated Central Asian glacial refugium, also shows a higher proportion of deleterious nsSNPs. The increased proportions of deleterious nsSNPs in both accessions may result from local adaptation to specific environmental conditions at the edge of the species range or the random fixation of slightly deleterious nsSNPs in small island populations due to genetic drift. The enrichment of deleterious nsSNPs is observed on a genome-wide level and thus is likely caused by higher levels of genetic drift. However, further studies are required to test whether the extensive local population structure in *A. thaliana* (Nordborg et al. 2005; Schmid et al. 2006), particularly in glacial refugia (Pico et al. 2008), contributes to variable levels of deleterious nsSNPs in response to differential selection and drift.

Consequences of domestication in rice

The rice sequences were used to examine differences between wild ancestors and domesticated cultivars. Both subspecies *O. sativa* ssp. *japonica* and *O. sativa* ssp. *indica* were probably domesticated independently in China from *O. rufipogon* (Kovach et al. 2007). For this reason, *O. sativa* and *O. rufipogon* cannot be grouped into separate monophyletic clades. Instead, *O. sativa* and Chinese *O. rufipogon* were combined into a single clade and compared to *O. rufipogon* accessions from a different geographic origin (Caicedo et al. 2007). Non-Chinese populations of *O. rufipogon* form a sister group of cultivated rice and the Chinese *O. rufipogon*. We observed fewer deleterious substitutions in wild than in cultivated rice. This result supports the hypothesis that domestication bottlenecks, artificial selection, and reduced purifying selection lead to an enrichment of deleterious amino acid substitutions, consistent with a previous study (Lu et al. 2006). Furthermore, intermediate frequency deleterious nsSNPs are enriched in *Oryza* suggesting a domestication effect, although this may be a statistical artifact of too few data points. Our simulations suggested that more data are necessary to reliably infer domestication effects on nsSNP distributions.

Application and comparison of nsSNP prediction programs

Plant genomes contain a high proportion of duplicated genes and the reliable inference of orthology and paralogy is crucial. It is better for nsSNP analysis to use few distant orthologous proteins than too many paralogous sequences

(Stone and Sidow 2005), because paralogs with a change in function may harbor different amino acids at critical positions. The use of proteins with altered function causes a decrease of sensitivity (Ng and Henikoff 2002; Stone and Sidow 2005) and paralogous proteins in the PSI-BLAST alignments might have considerably decreased our prediction accuracy.

By automating the alignment with PSI-BLAST, many sequences which overlap only in certain domains with the original protein were included in the alignment and remaining regions were filled with gaps that prevent the analysis of the corresponding regions. For this reason, the effects of eight substitutions in the *PHYB* protein could not be analyzed, although there was a correlation with certain phenotypes for three of them (Filiault et al. 2008). Generally, protein termini are difficult to align and were excluded from the prediction, but genome sequencing projects for numerous plant species will lead to improved sequence alignments and a higher prediction accuracy. The test data revealed a reasonable error rate of the predictions. Since false predictions can be assumed to occur uniformly across the genome, this observation suggests that the use of prediction programs is more appropriate for genome-wide analyses rather than inferring nsSNPs in individual genes.

The SIFT and MAPP prediction programs do not directly estimate the selection coefficients of nsSNPs. Prediction scores may be used as proxies for selection coefficients under the assumption that substitutions with higher scores have a stronger impact on protein function and fitness. Since a relationship between prediction scores and selection coefficients was not formally established, we used cutoff scores for a binary classification into tolerated and deleterious nsSNPs.

Detection of demographic effects by forward-in-time simulations

We used forward-in-time simulations to evaluate the effect of demographic history on the ratio of deleterious to tolerated nsSNPs and to estimate the power of the prediction programs to detect those effects. The total sequence length, i.e. the product of locus number, locus length, and numbers of individuals, was used as a first approximation to compare simulated and empirical datasets. The rice dataset had a total length of 0.8×10^6, the 2010 dataset 4.6×10^6, and the Perlegen dataset about 130×10^6 amino acids. In comparison, a simulation with 1,000 sampled individuals has a length of 6.6×10^6 amino acid residues in bottlenecked and non-bottlenecked populations combined. Such a dataset provided a power of more than 80% to detect differences in the ratio of deleterious to tolerated nsSNPs between the demographic histories of populations (Fig. 5). The rice and Perlegen data, but not the 2010 data revealed significant differences between groups of accessions. In contrast, the frequency distributions of deleterious and tolerated nsSNPs were significantly different in the 2010 and Perlegen, but not the rice dataset. The different results of the 2010 and Perlegen datasets are consistent with the simulations and show that whole genome resequencing is preferable to multilocus analysis for inferring the effects of demographic history on nsSNP frequencies.

Conclusions

Neutral genetic polymorphisms, such as silent SNPs are markers of choice for inferring the demographic history of a species. The present work shows that functional markers, such as nsSNPs can be utilized to infer the consequences of demographic history on the interplay of natural selection and genetic drift in plants. There is little power in the analysis of single genes, but genome-wide data carry enough information for comparisons of populations or species. NsSNPs are therefore useful markers for characterizing endangered plant species, plant genetic resources, or breeding populations. Since inbreeding and hitchhiking in response to artificial selection decrease genetic diversity, it will be interesting to infer deleterious nsSNP frequencies in germplasm of crop species to quantify fitness effects of plant domestication and modern breeding programs. Current genome sequencing projects of crop species and next generation sequencing technologies will greatly facilitate such investigations.

Acknowledgments We are grateful to the IPK bioinformatics group for assistance with the computer cluster and to two anonymous reviewers for their comments. This work was supported by core funding from the Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, and the Swedish University of Agricultural Sciences (SLU) Uppsala.

References

Abbott R, Gomes M (1989) Population structure and outcrossing rate of *Arabidopsis thaliana* (L) Heynh. Heredity 62:411–418

Albar L, Bangratz-Reyser M, Hebrard E, Ndjiondjop M, Jones M, Ghesquiere A (2006) Mutations in the *eIF(iso)4G* translation initiation factor confer high resistance of rice to *Rice yellow mottle virus*. Plant J 47:417–426

Alonso-Blanco C, de Vries HB, Hanhart CJ, Koornneef M (1999) Natural allelic variation at seed size loci in relation to other life history traits of *Arabidopsis thaliana*. Proc Natl Acad Sci USA 96(8):4710–4717

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Birney E, Clamp M, Durbin R (2004) Genewise and genomewise. Genome Res 14(5):988–995

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31(1):365–370

Brock M, Tiffin P, Weinig C (2007) Sequence diversity and haplotype associations with phenotypic responses to crowding: *GIGANTEA* affects fruit set in *Arabidopsis thaliana*. Mol Ecol 16(14):3050–3062

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, Bustamante CD, Purugganan MD (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3(9):e163

Cartegni L, Chew S, Krainer A (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3(4):285–298

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289–1303

Chisholm ST, Mahajan SK, Whitham SA, Yamamoto ML, Carrington JC (2000) Cloning of the *Arabidopsis RTM1* gene, which controls restriction of long-distance movement of tobacco etch virus. Proc Natl Acad Sci USA 97(1):489–494

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Scholkopf B, Nordborg M, Ratsch G, Ecker JR, Weigel D (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science 317(5836):338–342

Cronin JK, Bundock PC, Henry RJ, Nevo E (2007) Adaptive climatic molecular evolution in wild barley at the *Isa* defense locus. Proc Natl Acad Sci 104(8):2773–2778

El-Assal S, Alonso-Blanco C, Peeters A, Raz V, Koornneef M (2001) A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*. Nat Genet 29:435–440

Filiault DL, Wessinger CA, Dinneny JR, Lutes J, Borevitz JO, Weigel D, Chory J, Maloof JN (2008) Amino acid polymorphisms in *Arabidopsis* phytochrome B cause differential responses to light. Proc Natl Acad Sci 105(8):3157–3162

Friedman N, Ninio M, Pe'er I, Pupko T (2002) A structural EM algorithm for phylogenetic inference. J Comput Biol 9(2):331–353

Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci USA 99(2):1082–1087

Gazzani S, Gendall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. Plant Phys 132(2):1107–1114

Gepts P, Papa R (2002) Evolution during domestication. In: Encyclopedia of life sciences. Wiley, Chichester. http://www.els.net/

Hamblin MT, Casa AM, Sun H, Murray SC, Paterson AH, Aquadro CF, Kresovich S (2006) Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. Genetics 173(2):953–964

Hedrick P (1998) Maintenance of genetic polymorphism: spatial selection and self-fertilization. Am Nat 152(1):145–150

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

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci USA 101(29):10,667–10,672

Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C (2000) Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science 290(5490):344–347

Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47(6):713–719

Kimura M, Crow J (1963) The measurement of effective population number. Evolution 17(3):279–288

Kovach M, Sweeney M, McCouch S (2007) New insights into the history of rice domestication. Trends Genet 23:578–587

Lande R (1994) Risk of population extinction from fixation of new deleterious mutations. Evolution 48(5):1460–1469

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science 311(5769):1936–1939

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet 22(3):126–131

Maloof JN, Borevitz JO, Dabi T, Lutes J, Nehring RB, Redfern JL, Trainer GT, Wilson JM, Asami T, Berry CC, Weigel D, Chory J (2001) Natural variation in light sensitivity of *Arabidopsis*. Nat Genet 29(4):441–446

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37(9):997–1002

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11(5):863–874

Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12(3):436–446

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucl Acids Res 31(13):3812–3814

Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7(1):61–80

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3(7):e196

Oka H (1988) Origin of cultivated rice. Japan Scientific Societies Press, Tokyo, Elsevier, Amsterdam

Pico FX, Méndez-Vigo B, Martínez-Zapater JM, Alonso-Blanco C (2008) Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula. Genetics 180(2):1009–1021

Saitoh K, Onishi K, Mikami I, Thidar K, Sano Y (2004) Allelic diversification at the *C (OsC1)* locus of wild and cultivated rice: nucleotide changes associated with phenotypes. Genetics 168(2):997–1007

Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res 13:1250–1257

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169(3):1601–1615

Schmid KJ, Torjek O, Meyer R, Schmuths H, Hoffmann MH, Altmann T (2006) Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. Theor Appl Genet 112(6):1104–1114

Sekine KT, Ishihara T, Hase S, Kusano T, Shah J, Takahashi H (2006) Single amino acid alterations in *Arabidopsis thaliana RCY1* compromise resistance to *Cucumber mosaic virus*, but differentially

suppress hypersensitive response-like cell death. Plant Mol Biol 62(4):669–682

Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai YS, Gill BS, Faris JD (2006) Molecular characterization of the major wheat domestication gene *Q*. Genetics 172(1):547–555

Stein N, Perovic D, Kumlehn J, Pellio B, Stracke S, Streng S, Ordon F, Graner A (2005) The eukaryotic translation initiation factor 4E confers multiallelic recessive *Bymovirus resistance* in *Hordeum vulgare* (L.). Plant J 42(6):912–922

Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15(7):978–986

Suckow J, Markiewicz P, Kleina L, Miller J, Kisters-Woike B, Müller-Hill B (1996) Genetic studies of the *Lac Repressor XV*: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J Mol Biol 261(4):509–523

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–612

TIGR (2007) Rice genome annotation, vol 5. http://www.tigr.org/tdb/rice

Wolfe K, Li W, Sharp P (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84(24):9054

Wong GKS, Yang Z, Passey DA, Kibukawa M, Paddock M, Liu CR, Bolund L, Yu J (2003) A population threshold for functional polymorphisms. Genome Res 13(8):1873–1879

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. Science 308(5726):1310–1314

Yamasaki M, Wright S, McMullen M (2007) Genomic screening for artificial selection during domestication and improvement in maize. Ann Bot 100(5):967

Yeam I, Cavatorta JR, Ripoll DR, Kang BC, Jahn MM (2007) Functional dissection of naturally occurring amino acid substitutions in *eIF4E* that confers recessive potyvirus resistance in plants. Plant Cell 19(9):2913–2928