Molecular Biology and Evolution

Indel-Associated Mutation Rate Varies with Mating System in Flowering Plants

Jesse D. Hollister,*[,1] Jeffrey Ross-Ibarra,[2] and Brandon S. Gaut[1]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine

[2]Department of Plant Sciences, University of California, Davis

***Corresponding author:** E-mail: jhollister@oeb.harvard.edu.

Associate editor: Asger Hobolth

Abstract

A recently proposed mutational mechanism, indel-associated mutation (IDAM), posits that heterozygous insertions/deletions (indels) increase the point mutation rate at nearby nucleotides due to errors during meiosis. This mechanism could have especially dynamic consequences for the evolution of plant genomes, because the high degree of variation in the rate of self-fertilization among plant species causes differences in the heterozygosity of alleles, including indel alleles, segregating in plant species. In this study, we investigated the consequences of IDAM for species differing in mating system using both forward population genetic simulations and genomewide DNA resequencing data from *Arabidopsis thaliana*, *Oryza sativa*, and *Oryza rufipogon*. Simulations of different levels of selfing suggest that the effect of IDAM on surrounding nucleotide diversity should decrease with increasing selfing rate. Further simulations incorporating selfing rates and the time of onset of selfing suggest that the time since the switch to selfing also affects patterns of nucleotide diversity due to IDAM. Population genetic analyses of *A. thaliana* and *Oryza* DNA sequence data sets empirically confirmed our simulation results, revealing the strongest effect of IDAM in the outcrossing *O. rufipogon*, a weaker effect in the recently evolved selfer *O. sativa*, and the weakest effect in the relatively ancient selfer *A. thaliana*. These results support the novel idea that differences in life history, such as the level of selfing, can affect the per-individual mutation rate among species.

Key words: mutation rate, insertions/deletions, mating system, angiosperms.

Research article

Introduction

A key goal of research in population genetics and molecular evolution has been developing an understanding of the mechanisms generating genetic differences among organisms. Although natural selection, drift, and recombination each contribute to genetic differentiation, the variation upon which these forces act is ultimately provided by novel mutation. In particular, single nucleotide changes may underlie many of the phenotypic differences within and among species (Brookes 1999). However, other types of mutation, such as insertions/deletions (indels) and chromosome rearrangements, undoubtedly also contribute to phenotypic variation (Haase et al. 2008; McCarroll et al. 2008).

Several mechanisms are thought to produce single-nucleotide mutations, including replication errors during DNA synthesis, spontaneous DNA lesions, and production of mutagenic substrate nucleotides (Maki 2002). In addition, there is a growing body of evidence to support the mutagenicity of meiotic recombination (Gaut et al. 2007). Adding to this array of known mechanisms, Tian et al. (2008) recently proposed a novel pathway for the generation of single-nucleotide mutations: indel-associated mutation (IDAM). Under this model, which causally links indels and nucleotide substitutions, heterozygosity for the presence of an indel increases the single-nucleotide mutation rate at closely linked sites during meiosis. Using multispecies alignments of orthologous genomic regions, Tian et al. (2008) demonstrated greater nucleotide divergence in regions closely linked to indels and showed that this effect decreases with increasing distance from an indel. Using polymorphism data from three sequenced lines of yeast, they also estimated that the presence of segregating indels may increase mutation rates by as much as 35-fold within ~200 bp of the indel.

In their analysis, Tian et al. (2008) dealt primarily with outbreeding organisms. They pointed out that, in random-mating populations, a novel indel mutation destined to rise to fixation will spend 50% of its time in heterozygotes. However, inbreeding changes the expected genotype frequencies of individuals within a population, leading to a paucity of heterozygous individuals. Wright (1921) defined the equilibrium-inbreeding coefficient F to describe the differences in genotype frequencies in inbreeding populations. Using F as a measure of inbreeding, the time a novel indel allele spends in heterozygous state en route to fixation is $(1 - F)/2$ (see Supplementary Materials online). Clearly, as populations approach complete inbreeding ($F = 1$), indel alleles fixed in the population will have spent virtually no time in heterozygotes.

Because selfing reduces the frequency of heterozygous individuals, it has several well-characterized effects. For example, recessive mutations with beneficial effects are more likely to fix in inbreeding populations, as are chromosomal rearrangements with deleterious effects on heterozygote fitness. Inbreeding also decreases the effective population size, increasing the likelihood that mildly deleterious alleles

will go to fixation, although this effect is small (Charlesworth 1992). The discovery of the IDAM mechanism suggests that selfing could also strongly influence mutation rates near heterozygous indels and thus affect patterns of nucleotide diversity.

Flowering plants vary widely in rates of self-fertilization (Vogler and Kalisz 2001), from highly outcrossing species with elaborate self-incompatibility mechanisms to well-known self-fertilizing species, such as *Arabidopsis thaliana* and *Oryza sativa*, that have rates of inbreeding 90% or higher. In this paper, we take advantage of this diversity to investigate the implications of mating system for the IDAM model. We first utilize computer simulation to obtain an expectation of the joint effects of inbreeding and IDAM on nucleotide variation. We then examine a genomewide polymorphism data set from *A. thaliana*, in order to test predictions from our simulations of IDAM in a highly selfing species. Finally, we simulate the interplay between IDAM, selfing rate, and time since the evolution of selfing and compare simulation results with DNA sequence data from *A. thaliana*, *O. sativa*, and *Oryza rufipogon*.

Materials and Methods

Forward Simulation

To investigate the effects of inbreeding on indel-associated diversity, we constructed a two-locus forward population genetic model. The model posits an indel locus evolving under an infinite-allele mutation model completely linked to a second sequence locus evolving under an infinite-site model. Individuals are diploid, and populations evolve forward in time under the standard Wright–Fisher model. In addition to the standard parameters of population size N and mutation rate μ (μ_i for the indel locus and μ_s for linked sites) we include three parameters: 1) a scalar R that modifies the μ_s in individuals heterozygous for alleles at the indel locus, 2) the probability S of self-fertilization (where $S = 2F/(1 + F)$) (Hedrick 2005), and 3) the timing T of the evolution of self-compatibility. Simulations were run for $12N$ generations before data were collected to ensure that equilibrium conditions had been reached in the simulated populations. Each simulation begins with a monomorphic population and no selfing ($S = 0$); after $12N - T$ generations inbreeding changes instantaneously to $S > 0$. Levels of heterozygosity and the site-frequency spectrum (SFS) of both the indel and sequence loci were compared with analytical expectations for $R = 1$ and $T = 0$ to ensure accuracy of the simulations.

Sequence Data and Analysis

We made use of two large DNA sequence data sets for this study (Nordborg et al. 2005; Caicedo et al. 2007). Nordborg et al. (2005) resequenced 876 fragments of approximately 500–800 bp in a panel of 96 *A. thaliana* accessions (supplementary table 1, Supplementary Material online). Caicedo et al. (2007) resequenced 111 fragments, 400–550 bp in length, in several taxa of the genus *Oryza*. These

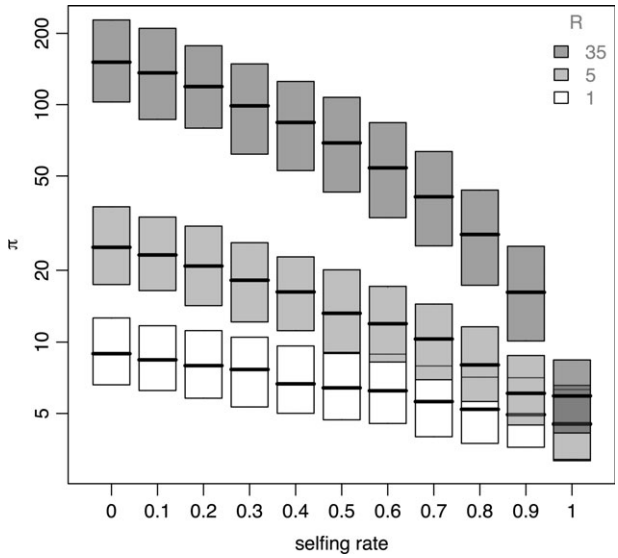

FIG. 1. Nucleotide diversity (π) for three relative rates of indel-induced mutation (R) in populations with varying rates of self-fertilization. Shown is the median (bar) and interquartile range (box) of total π linked to both indel and nonindel loci from 1,000 simulations. Simulations assume $\theta = 4N\mu_i = 1$ for indel mutations and $\theta = 4N\mu_s = 10$ for linked variation in individuals homozygous at the indel locus. $R = 1$ shows the effect of selfing on π in the absence of indel effects: As predicted by theory, π drops to 50% of the random-mating expectation as the selfing rate increases to 1. Higher values of R greatly increase the differences in π due to differences in selfing rate. Note the log scale on the *y*-axis.

fragments were first aligned using ClustalX (Thompson et al. 1997), and the resulting alignments were confirmed by hand, recoding sequence end gaps as missing data to avoid their conflation with indel alleles. From these alignments, we included 46 *Oryza sativa* spp. *japonica*, 26 *O. sativa* spp. *indica*, and 21 *O. rufipogon* individuals for analysis (supplementary table 2, Supplementary Material online).

All analysis of variation made use of a folded SFS, considering only the frequency of the minor allele. Polymorphism patterns for all loci were analyzed using custom perl scripts and software from the analysis package of the libsequence C++ library (Thornton 2003).

Results

Simulation of IDAM: The Effect of Selfing

To investigate the interaction between the rate of selfing and the effect of IDAM on patterns of single nucleotide polymorphism (SNP) diversity, we simulated variation at sites linked to an indel under a simple two-locus Wright–Fisher model with no recombination. We simulated data across a grid of probabilities of self-fertilization (S) for each of three relative rates of IDAM. Figure 1 summarizes simulated data for $R = 1$, 5, and 35, with $\theta_i = 4N\mu_i = 1$ for indel mutations and $\theta_s = 4N\mu_s = 10$, but results with other parameter values were qualitatively similar (data not shown). These data reveal three fundamental features of

Table 1. Summary of Data Used in This Paper.

	n	n_i	Median π_i	Median π_o	PVE Observed (%)	PVE Simulated (%)
Oryza rufipogon	111	41	0.005	0.0019	21	11.6
Oryza sativa indica	111	31	0.0018	0.001	11	7
Oryza sativa japonica	111	41	0.0019	0.00028	10	6.5
Arabidopsis thaliana	876	544	0.003	0.0028	7	1.3

Columns represent number of sequenced regions (n), number of regions with indels (n_i), nucleotide diversity per base pair for regions with (π_i) and without (π_o) indels, and percent variance in diversity explained (PVE) by presence of indels in observed data and for simulated data sets.

the IDAM model with respect to the selfing rate (S). First, there is a strong negative correlation between self-fertilization and diversity over the range of R values investigated (Spearman's *rho* = −0.38, −0.59, and −0.66 for R = 1, 5, and 35, respectively). Second, the strength of the negative correlation is strongly dependent on R. When R = 1 (no IDAM), the increase in selfing rate alone is responsible for the decrease in diversity. In this case, there is a 2-fold reduction in diversity as F goes from 0 to 1, as predicted by theory (Nordborg 2000). However, when R = 5 or 35, there is a 6-fold or 25-fold reduction in diversity, respectively, representing the interplay between IDAM and inbreeding: The mutation rate at nucleotides linked to heterozygous indels increases with R, whereas the number of generations during which indels are heterozygous decreases with F. Finally, even with a low R value (R = 5) and high inbreeding (F = 0.90), there is a detectable IDAM effect (median π/locus = 4.9 and 6.2 for R = 1 and 5, respectively; Mann–Whitney U Test, $P < 2.2e-16$). These simulations clearly illustrate that the strength of the IDAM effect depends on F, but they also suggest that the effect should be detectable even for very high values of inbreeding (supplementary fig. S1, Supplementary Material online).

IDAM in *A. thaliana*

Our simulations predicted a weak but discernable effect of IDAM on SNP diversity in a selfing species (fig. 1). *Arabidopsis thaliana* reproduces almost exclusively via self-fertilization (Abbott and Gomes 1989), a mode of reproduction that likely evolved a million or more years before the present (ybp) (Tang et al. 2007). Heterozygous indel variants are likely rare in *A. thaliana* individuals, and it is unlikely that presently segregating variation has persisted from prior to the evolution of self-compatibility in this species. The IDAM model posits increased mutation rates only in heterozygous individuals, but our simulation results suggest that IDAM effects may be detectable in *A. thaliana* nonetheless.

We examined patterns of nucleotide and indel diversity in 876 sequenced fragments spaced approximately every 100 kb along all five chromosomes of the *A. thaliana* genome (Nordborg et al. 2005). We compared nucleotide diversity (π/site) for all sites in fragments with (n = 544) or without (n = 332) indels (table 1). Presence of at least one indel significantly impacted total nucleotide diversity (π) ($P < 1.5e-14$), and explained ~7% of the total variance in π. Moreover, π increased as a function of the number of indels per sequenced fragment (r^2 = 0.24, $P < 2.2e-16$) and decreased as a function of increasing distance from indels in fragments containing indels (fig. 2; Spearman's rho = −0.21, $P < 2.2e-16$). These patterns are consistent with those observed in Tian et al. (2008) and indicate an association between segregating indels and increased nucleotide variation in *A. thaliana*.

It is possible, however, that such an association could arise from differences in constraint among genomic regions: Low-constraint regions may simply harbor more mutations, including both indels and segregating nucleotides. To address this issue, we examined nucleotide diversity at synonymous sites in fragments with or without indels. With the possible exception of weak selection for codon usage, these sites likely evolve neutrally, thus allowing examination of the effect of indel variation essentially independent of constraint. Variation at synonymous sites was significantly higher for fragments containing indels (linear model r^2 = 0.014, $P < 0.0005$), and thus, sites that are subject to similar levels of constraint differ in levels of variation in a manner consistent with IDAM.

Another method for assessing constraint among protein-coding regions is to compare the SFS of synonymous and nonsynonymous segregating sites in those regions. Constraint on protein-coding regions is expected to skew the SFS of nonsynonymous variants toward low-frequency values compared with synonymous variants, due to weak purifying selection on mildly deleterious nonsynonymous changes (Fay et al. 2001). If constraint is lowered, but still present, a larger proportion of mildly deleterious alleles are expected to segregate. If constraint is lower in indel-containing protein-coding fragments, we thus predict that they should also harbor an excess of low-frequency nonsynonymous mutations relative to synonymous sites when compared with regions lacking indels. However, the ratios of low-frequency (<10%) nonsynonymous to synonymous changes were similar in regions with and without indels, indicating no consistent differences in constraint between such regions (Fisher's exact test P = 0.45).

The IDAM model predicts that, as an indel drifts to high frequency in a population, closely linked sites will accumulate single-nucleotide mutations due to mutagenicity of the indel. These indel-induced mutations occur in addition to other single-nucleotide mutations occurring by mechanisms unrelated to presence of the indel. This should result in a proportionally higher level of SNP variation near high-frequency indels than near nonmutagenic alleles at the same frequency (fig. 3A). To test this hypothesis, we measured nucleotide variation in regions containing a

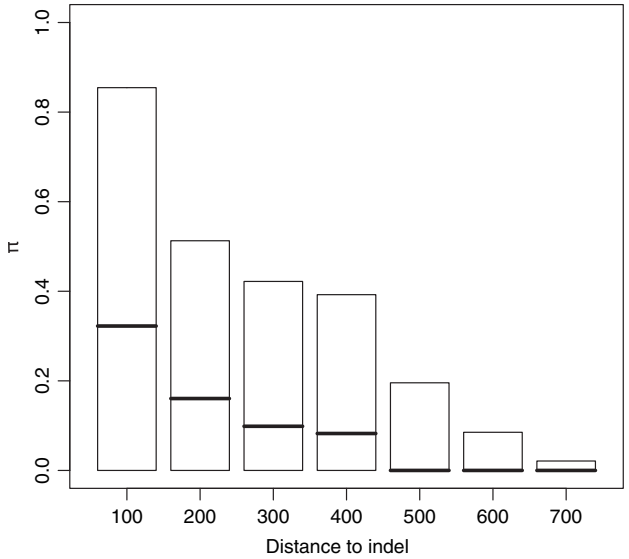

FIG. 2. Nucleotide diversity, π/base pair, for 100-bp bins of increasing distance from the nearest indel in the *Arabidopsis thaliana* DNA sequence data set. Shown is the median (bar) and interquartile range (box) for each bin. Nucleotide diversity decreases with increasing distance from the nearest indel.

randomly chosen indel or SNP in the sample of *A. thaliana* sequenced fragments. SNPs were selected from fragments regardless of whether the region also harbored indels. We then separately considered fragments in which the indel or SNP was at low (minor allele) frequency ($<$0.25), or high frequency ($\geq$0.25). We reasoned that, under the IDAM model, the ratio of the level of variation around high-frequency indels to the level of variation around high frequency SNPs should be greater than the equivalent ratio for low-frequency indels and SNPs. Consistent with this prediction, high-frequency indels exhibit proportionally more SNP variation at linked sites than high-frequency SNPs. The ratio of median π/site in fragments harboring high-frequency indels versus high-frequency SNPs is 2.7:1, whereas that ratio for low-frequency indels versus SNPs is 1.3:1 (fig. 3*B*). To test the significance of the difference in ratios, we performed 10,000 permutations of π values among the four frequency-mutation classes and found that the ratios of median π were never as different as in the observed data ($P < 1e-4$). Taken together with the evidence against differences in constraint between indel and non–indel-containing regions, these results are consistent with a detectable effect of IDAM in *A. thaliana*.

Simulation of IDAM in *A. thaliana*, *O. sativa*, and *O. rufipogon*

Next, we performed simulations to assess how IDAM effects should vary across taxa, based upon differences in selfing rate and timing of the switch to selfing. The simulations used S and T values that mimic estimates of these parameters for *A. thaliana* and *Oryza*. For example, the selfing rate for *A. thaliana* was assumed to be $S = 0.99$ (Abbott and Gomes 1989), and the timing to the shift of selfing was assumed to be $T = 8N$ (Tang et al. 2007). Similarly, $S = 0.99$ for *O. sativa*, but $S = 0.68$ for *O. rufipogon* (Gao et al. 2007). The timing of the evolution of selfing for *O. sativa* was assumed to be $T = 0.3N$, and S for *O. rufipogon* was held constant for the entire simulation (e.g., $T = 12N$), consistent with the inference of mixed mating as the ancestral state in *Oryza* (Oka 1988). Simulations sampled the same number of loci as the observed data and were checked to ensure that overall levels of nucleotide diversity

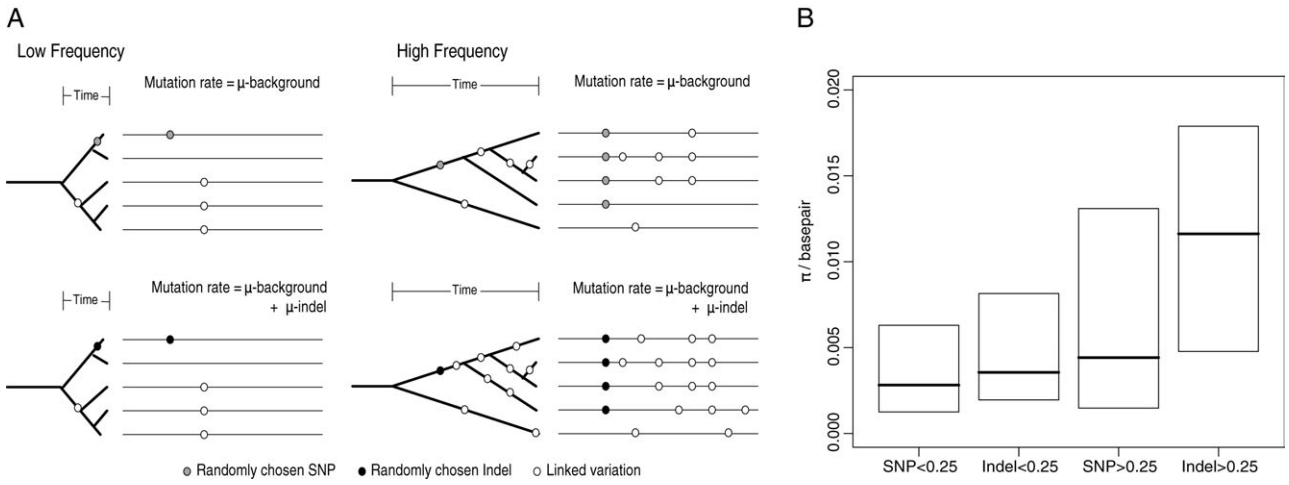

FIG. 3. *A*) Illustration of the expectation for levels of variation surrounding randomly chosen indels and SNPs. Trees represent gene genealogies and horizontal lines represent chromosome sequences for low-frequency (left) or high-frequency SNPs or indels. Under the IDAM model, regions containing high-frequency indels should accumulate proportionally more variation than regions containing high-frequency SNPs due to mutagenicity of the indel. *B*) Levels of linked variation around randomly chosen low ($<$0.25) and high ($>$0.25) frequency indels and SNPs. Shown is the median (bar) and interquartile range (box) for each group. The ratio of linked variation around high-frequency indels compared with high-frequency SNPs is proportionally greater than the ratio around low-frequency indels compared with low-frequency SNPs, consistent with an elevated mutation rate due to IDAM.

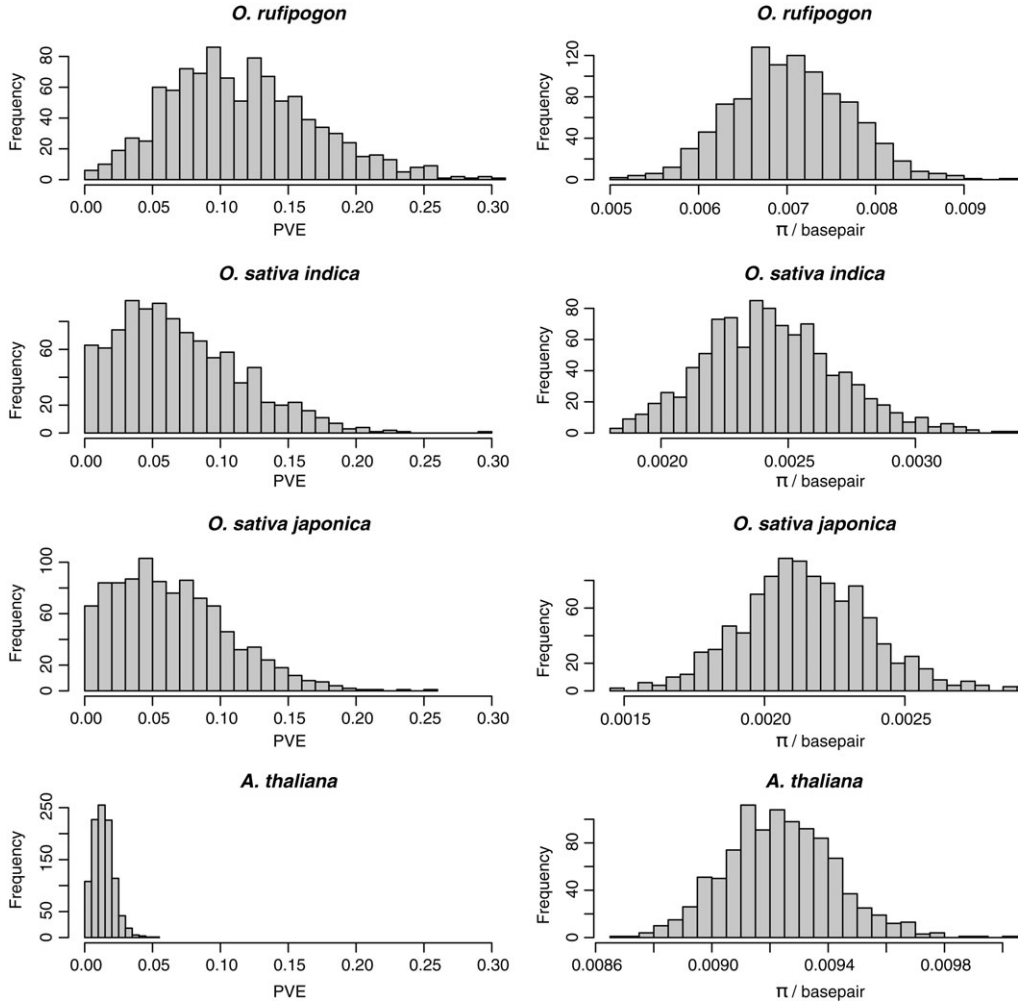

FIG. 4. Results from forward genetic simulation of indel-associated mutation with levels of selfing representative of *Oryza rufipogon*, *Oryza sativa* ssp. *indica* and *japonica*, and *Arabidopsis thaliana*. The left column shows the distribution of percent variance in SNP diversity explained by presence of closely linked indels (PVE) for the four sets of 1,000 simulations. The right column shows the distribution of π/base pair over all simulations, calculated by dividing the π/locus from each simulation by the average length of sequence segments in the observed data.

were similar to our observed data (table 1). *R* was set to 35 for all simulations, reflecting estimates from Tian et al. (2008). From simulations using these parameter values, we calculated the mean percentage of variance in nucleotide diversity (π) explained by the presence of indels from 1,000 simulated multilocus data sets (fig. 4).

The simulation results suggest that when observed levels of diversity and estimated selfing rates are taken into account, *O. rufipogon* should display the greatest effect of indel presence/absence on the variance in π, due to a high outcrossing rate and thus a greater proportion of heterozygous indels. The mean percentage of variance in π explained by indel presence/absence was 11.6% across simulations for *O. rufipogon*. The simulations of the domesticated ssp. of *O. sativa*, with much lower rates of outcrossing (1%) and recently evolved self-compatibility, predicted 7% and 6.5% of the variance in π explained by indels for spp. *indica* and ssp. *japonica*, respectively. For *A. thaliana*, with an outcrossing rate of 1% and a long history of selfing (switch to selfing ~1 Ma), the simulations predicted only 1.3% of the variance in π explained by indel presence/absence. Thus, the simulations showed an effect of both the level of outcrossing and the time since the evolution of self-compatibility on the percent variance in π explained by the presence of indels. Moreover, these results provided a theoretical basis for comparison to empirical data.

IDAM in *Oryza*

To compare our simulation results with potential differences in IDAM among closely related taxa with different mating systems, we next investigated patterns of SNP and indel variation in 111 expressed sequence tag (EST)–based sequence fragments in two subspecies of domesticated rice (*O. sativa* spp. *japonica* and *indica*), as well as their wild ancestor *O. rufipogon* (Caicedo et al. 2007). Both *japonica* and *indica* are highly selfing, with outcrossing rates estimated at 1%, similar to the rate in *A. thaliana*. Unlike *A. thaliana*, however, self-fertilization evolved comparatively recently in domesticated rice, probably concurrent with domestication approximately 10,000 ybp. *Oryza*

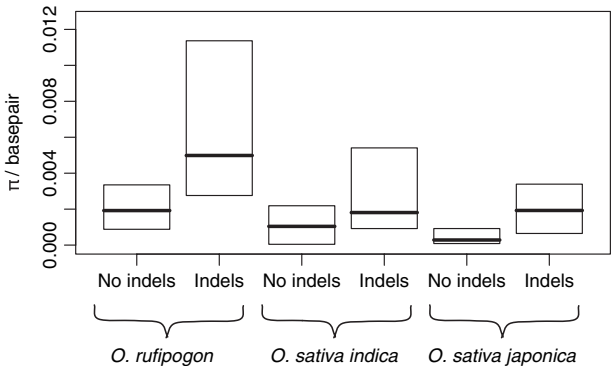

Fig. 5. Nucleotide diversity in sequenced fragments with and without indels for *Oryza rufipogon* and two subspecies of *Oryza sativa*. Shown is the median (bar) and interquartile range (box) for each group. Presence of at least one indel explains roughly twice the variance in SNP diversity in *O. rufipogon* compared with *O. sativa*.

rufipogon, considered the probable ancestor of domesticated rice, has a mixed-mating system, with estimates of outcrossing rates of 30–52% (Gao et al. 2007). On the basis of our simulation results (figs. 1 and 4), we predicted that all three taxa would exhibit a greater effect of IDAM than *A. thaliana*, with the clearest association being observable in *O. rufipogon*.

Consistent with our observations for *A. thaliana*, all three taxa showed higher total nucleotide variability in fragments with at least one indel (table 1; fig. 5). Also consistent with our simulation results, *O. rufipogon* had the strongest effect of indel presence on variation within a sequenced fragment ($r^2 = 0.21$, $P < 4e-7$). *Oryza sativa* ssp. *indica* and *japonica* showed a smaller effect ($r^2 = 0.11$, $P < 0.0005$ and $r^2 = 0.10$, $P < 0.0007$, respectively), but the effect was still greater than in *A. thaliana* ($r^2 = 0.07$, $P < 1.5e-14$). Indel-associated variation thus appears to be correlated with published estimates of the rate and time since the evolution of selfing in *Oryza* and *Arabidopsis*.

To more rigorously quantify the interaction between selfing and IDAM, we combined the *Oryza* data into a single data set. Using the combined data set, we constructed a linear mixed model that included mating system as a fixed effect, with taxonomic status, indel presence, and the interaction of indel presence with mating system as random effects. This analysis showed that almost as large a percentage of the variance (15.9%) was explained by the interaction of mating system with indel presence as by indel presence alone (18.2%). This result underscores the importance of mating system in determining the effects of IDAM on genetic diversity.

Discussion

In a recent study, Tian et al. (2008) used six genomic comparisons, with outgroups, to investigate SNP divergence around indels. They showed that SNP divergence 1) is elevated near indels, 2) decreases as a function of distance to indel, and 3) is positively correlated with the abundance of indels in genomic regions. To explain these patterns, they proposed the IDAM model, which posits that heterozygous indels increase the single-nucleotide mutation rate at closely linked sites due to errors during meiosis. This study showed convincing evidence of an effect of IDAM on the rate of divergence in multiple species comparisons. However, thus far there has been little evidence for the model provided by population-level data. Because this model invokes a population genetic process (higher mutation rates near heterozygous indels), we made use of forward genetic simulations and several large population-level DNA sequence data sets to evaluate the interaction of IDAM and selfing on patterns of diversity in plants.

Our simulation results suggested that selfing reduces the magnitude of IDAM and that the effect of IDAM should depend both on the rate of selfing and the time since evolution of self-fertilization in a lineage. Consistent with these predictions, we observed the strongest associations between indel presence and π in the mixed-mating *O. rufipogon*, an intermediate association in the recently evolved selfing lineages of *O. sativa*, and a weaker, but still significant association in the ancient selfer *A. thaliana*. Thus, the empirical data are consistent with simulations showing an interaction between IDAM and mating system.

Our simulations predicted a weaker effect of IDAM for all three species than the effect we detected in the sequence data. At least three factors could account for this underestimate: First, we relied on single published estimates of the selfing rate for these species and for the evolution of selfing in *A. thaliana* and *O. sativa*. If the selfing rates in these populations vary, we may have systematically overestimated the prevalence of selfing in our simulated populations. The difference in magnitude between the simulated and empirical results was greatest in *A. thaliana*. In our simulations, we set $T = 8N$ (~1 Ma) for *A. thaliana* (Tang et al. 2007). If the actual timing of the switch to selfing is significantly less for this species, individuals may still harbor segregating variation that arose prior to the evolution of selfing, which could account for the difference in percent variance explained between the simulated and empirical data for *A. thaliana*. Second, the parameter *R*, the increase in mutation rate near indels, may be higher than estimated in Tian et al. (2008), possibly leading to an underestimate of the IDAM rate. Third, artifacts in the sequence alignments around indel breakpoints may have inflated our estimates of diversity in these regions. However, although indels can be difficult to resolve using many alignment algorithms (Lunter et al. 2008; Cartwright 2009), we estimated total nucleotide variability for each fragment across the entire region (~500 bp), making it unlikely that alignment uncertainty around indels accounts for a large proportion of the levels of variation we inferred.

A probable mechanism underlying IDAM is mutagenesis during meiotic chromosome pairing (Tian et al. 2008). Our simulations use the parameter, *R*, to approximate the effect of heterozygous indels on the mutation rate at nearby nucleotides. However, we did not explicitly incorporate

recombination between indel and SNP loci in our simulations, for two reasons. First, the simulations were specifically tailored for comparison with resequencing data from *A. thaliana* and *Oryza*, and these sequenced regions are quite small (~500 bp), with little evidence of intralocus recombination (data not shown). Genomewide, linkage disequilibrium is thought to extend over 10 kb in the *A. thaliana* genome (Kim et al. 2007) and even farther in *Oryza* (75–150 kb in *O. sativa*; 40 kb in *O. rufipogon*) (Mather et al. 2007). Second, and more importantly, recombination will not affect the patterns of polymorphism within a given sequenced region; it will only affect how polymorphisms are partitioned among haplotypes. Because we examined such patterns among regions in randomly chosen individuals both possessing and lacking indels, the patterns we observed in the resequencing data are directly comparable to the simulation results.

In the *Oryza* resequencing data, we cannot rule out the possibility that a combination of decreased constraint in regions bearing indels and greater effective population size in *O. rufipogon* contribute to the stronger association between indels and increased SNP variation in this species compared with *O. sativa*. However, we favor the interpretation that a combination of IDAM and selfing rate is driving this pattern, for two reasons. First, all regions selected for resequencing in *Oryza* were chosen based on homology with EST data and primers were designed within exons for all fragments (Caicedo et al. 2007), so there is no a priori reason to assume that constraint differs markedly among these regions. Second, the hypothesis that differences in constraint and effective populations size account for this pattern is not consistent with the differences in variation between the domesticated subspecies of *O. sativa*. In ssp. *indica*, the median π/bp is twice as high as in ssp. *japonica* (0.0012 vs. 0.00056), indicating a higher effective population size, but the association between indels and increased SNP variation is nearly identical in both (11% vs. 10%). This suggests that an interaction between selfing rate and IDAM, rather than a combination of effective population size and constraint, has shaped patterns of SNP variation around indels in *Oryza*.

This study reveals that the rate of selfing may affect mutation rates near indels. This finding has important repercussions for understanding the generation of variation within populations and divergence among species. Selfing is known to decrease within-population variation because it reduces the effective population size (Charlesworth and Wright 2001; Glemin 2007). Selfing also increases the likelihood of fixation of beneficial recessive mutations as well as mildly deleterious mutations, although the latter effect may be small (Charlesworth 1992). However, selfing also "unmasks" deleterious recessive alleles, promoting their preferential loss (Barrett and Charlesworth 1991), but it is unclear to what extent this unmasking effect balances the greater chance of fixation of such alleles in selfing populations. This study adds to the array of known effects of selfing on the rate of evolution, providing evidence that the rate of selfing also influences the per-individual mutation rate via differences in the mutagenic effect of segregating indels. Thus, highly selfing species may exhibit a higher rate of evolution via increased fixation of beneficial recessive alleles and mildly deleterious alleles but a lower rate of neutral evolution and less segregating variation around indels due to decreased IDAM.

An important corollary of this concerns the rate of divergence between species. A long-held tenet of population genetic theory states that the rate of fixation of neutral mutations, k, is equal to μ, the per-individual mutation rate; that is, the rate of substitution is independent of population-level parameters (Kimura 1983). If our interpretation of the evidence is correct, then the per-individual rate of single-nucleotide mutations near indels, and therefore the rate of nucleotide substitution in these regions, is influenced by the rate of inbreeding. This has implications for constancy of the molecular clock among taxa, especially those that vary widely in mating system, such as flowering plants.

Our results raise additional considerations surrounding genome-scale patterns of variation. Differences in evolutionary rate driven by IDAM depend not only on the rate of selfing and R but also on the density of indels and their average fitness effects. Tian et al. (2008) point out that, because indels are more prevalent in noncoding DNA, their effect on the rate of phenotypic evolution may be biased toward changes in gene expression via cis-regulatory mutations. In any case, we predict that the rate of sequence evolution will have the greatest variance among related selfing and outcrossing species in intergenic regions, because these regions have a higher number of segregating indels.

In conclusion, our study uncovered strong support for IDAMs contributing to polymorphism patterns in *A. thaliana* and *Oryza*. The magnitude of the effect of indels correlates inversely with the rate of inbreeding in these plant species, and our simulations of IDAM support the notion that these differences are due to the relationship between selfing and heterozygosity. As the selfing rate increases, newly arising indels spend fewer generations in heterozygotes and therefore contribute less to increases in mutation rates for surrounding nucleotides. Thus, the rate of neutral evolution among species cannot be considered strictly independent of population-level processes.

Supplementary Material

Supplementary figure S1 and supplementary tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to thank G. Coop, K. Thornton, and 3 anonymous reviewers for valuable suggestions concerning an earlier version of this manuscript.

References

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

Barrett SC, Charlesworth D. 1991. Effects of a change in the level of inbreeding on the genetic load. *Nature* 352:522–524.
Brookes AJ. 1999. The essence of SNPs. *Gene*. 234:177–186.
Caicedo AL, Williamson SH, Hernandez RD, et al. (12 co-authors). 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet*. 3:1745–1756.
Cartwright RA. 2009. Problems and solutions for estimating indel rates and length distributions. *Mol Biol Evol*. 26:473–480.
Charlesworth B. 1992. Evolutionary rates in partially self-fertilizing species. *Am Nat*. 140:126–148.
Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Curr Opin Genet Dev*. 11:685–690.
Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics*. 158:1227–1234.
Gao H, Williamson S, Bustamante CD. 2007. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 176:1635–1651.
Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet*. 8:77–84.
Glemin S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* 177:905–916.
Haase B, Jude R, Brooks SA, Leeb T. 2008. An equine chromosome 3 inversion is associated with the tobiano spotting pattern in German horse breeds. *Anim Genet*. 39:306–309.
Hedrick P. 2005. Genetics of populations. Sudbury (MA): Jones and Bartlett.
Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet*. 39:1151–1155.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge (UK): Cambridge University Press. xv, p. 367.
Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res*. 18:298–309.
Maki H. 2002. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annu Rev Genet*. 36:279–303.
Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. 2007. The extent of linkage disequilibrium in rice (*Oryza sativa L.*). *Genetics* 177:2223–2232.
McCarroll SA, Huett A, Kuballa P, et al. (17 co-authors). 2008. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet*. 40:1107–1112.
Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics*. 154:923–929.
Nordborg M, Hu TT, Ishino Y, et al. (24 co-authors). 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol*. 3:e196.
Oka HI. 1988. Origin of cultivated rice. Japan Sci Soc Press/Elsevier, Tokyo (Japan)/Amsterdam (Netherlands):.
Tang C, Toomajian C, Sherman-Broyles S, Plagnol V, Guo YL, Hu TT, Clark RM, Nasrallah JB, Weigel D, Nordborg M. 2007. The evolution of selfing in *Arabidopsis thaliana*. *Science* 317: 1070–1072.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 25:4876–4882.
Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.
Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455:105–108.
Vogler DW, Kalisz S. 2001. Sex among the flowers: the distribution of plant mating systems. *Evolution* 55:202–204.
Wright S. 1921. Systems of mating. II. the effects of inbreeding on the genetic composition of a population. *Genetics* 6:124–143.