Hitchhiking Selection Is Driving Intron Gain in a Pathogenic Fungus

Patrick C. Brunner,*[1] Stefano F.F. Torriani,[1] Daniel Croll,[1] Eva H. Stukenbrock,[2] and Bruce A. McDonald[1]

[1]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

[2]Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

***Corresponding author:** E-mail: patrick.brunner@usys.ethz.ch.

Associate editor: Jianzhi Zhang

Abstract

The variability of intron density among eukaryotes is puzzling and still debated. Most previous studies have been limited because of the near absence of intron presence–absence polymorphism (IPAP) within species or because comparisons could be made only between distantly related species. We conducted population genetic analyses on eight loci showing IPAP to investigate the effect of natural selection on intron dynamics in a global collection of the panmictic fungal plant pathogen *Zymoseptoria tritici* and its very close relatives. Five of these loci likely represent recent intron gains because their absence is fixed among the closest relatives of *Z. tritici*, and three likely represent recent intron losses because their presence is fixed among the close relatives. We analyzed signatures of selection by comparing allele frequencies, nucleotide diversities, and rates of recombination and found compelling evidence that at least two out of the five intron-gain loci, a SWIM zinc-finger gene and a sugar transporter, are under directional selection favoring alleles that gained the intron. Our results suggest that the intron-present alleles of these loci are sweeping to fixation, providing a genetic hitchhiking mechanism to explain rapid intron gain in *Z. tritici*. The overall findings are consistent with the hypothesis that intron gains are more likely to be driven by selection while intron losses are more likely to be due to neutral processes such as genetic drift.

***Key words:* spliceosomal introns, *Zymoseptoria tritici*, genetic hitchhiking, *Mycosphaerella graminicola*, natural selection, genetic drift.**

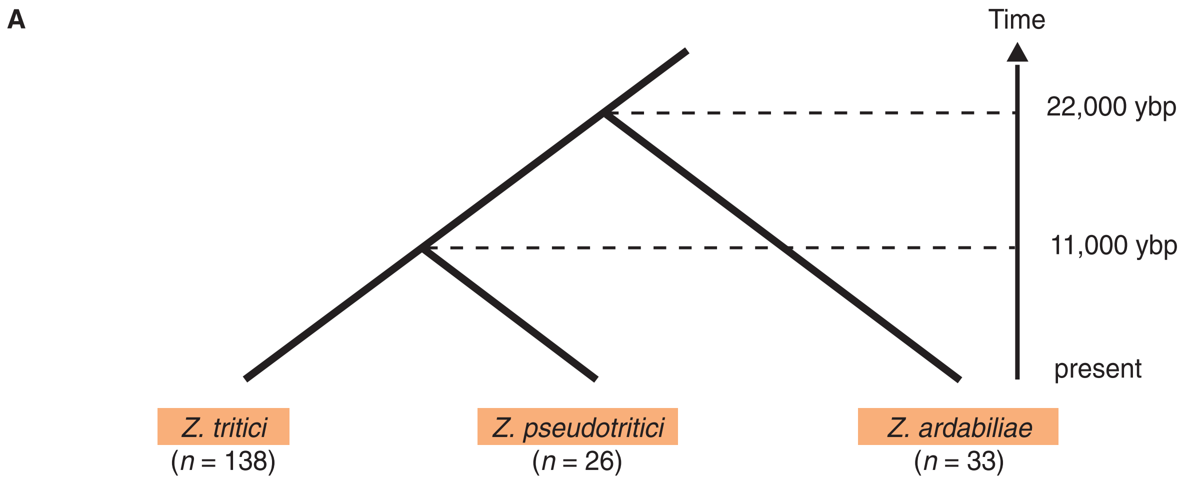
Introduction

Though spliceosomal introns are a major structural component of most eukaryotic genes, and intron density varies by more than three orders of magnitude among eukaryotes, the source of this puzzling variation as well as the underlying evolutionary forces responsible for intron gains or losses are still debated (reviewed by Jeffares et al. 2006; Roy and Gilbert 2006; Lynch 2007). Using population genetic theory, Lynch (2002) showed that intron evolution can be viewed as a population genetic process in which intron loss or gain is determined by evolutionary mechanisms such as natural selection or genetic drift. Evolutionary analyses based on comparing genetic variation within and between species provide powerful tools to investigate the role of natural selection and genetic drift on intron dynamics. However, the near absence of intron presence–absence polymorphism (IPAP) has so far limited this kind of approach, with two exceptions investigated in *Drosophila teissieri* (Llopart et al. 2002; Li et al. 2009) and *Daphnia pulex* (Omilian et al. 2008).

In a recent study, we utilized population genomic comparisons to identify transitory phases of intron gains and losses in three closely related fungi (Torriani et al. 2011) and identified 74 intron positions showing intraspecific presence–absence polymorphisms for the entire intron. *Zymoseptoria tritici* (syn. *Mycosphaerella graminicola*) is a globally distributed plant pathogenic fungus (Eyal 1999) that causes the most economically damaging foliar disease of wheat in Europe (Orton et al. 2011). *Zymoseptoria tritici* is characterized by a high effective population size, regular recombination, and significant gene flow (Zhan et al. 2003). *Zymoseptoria tritici* has two closely related sister species, *Z. pseudotritici* and *Z. ardabiliae*, that were recently discovered infecting uncultivated grasses in the Middle East. It was hypothesized that *Z. tritici* emerged as a pathogen specialized to infect wheat during the domestication of wheat in the Fertile Crescent approximately 10,000 years ago (Stukenbrock et al. 2007). This recently emerged host–pathogen system provides a rare opportunity to investigate the evolutionary processes shaping the genome of an emerging pathogen.

For this study, we randomly selected eight of the previously detected loci showing IPAPs in *Z. tritici* (fig. 1) for a more detailed analysis. All introns were described in Torriani et al. (2011). Five out of the eight IPAPs belong to one of the 38 identified intron families of *Z. tritici*. These introns most likely were involved in intraspecific intron transfer or transposition. Five of these loci did not contain the intron in the sister species *Z. pseudotritici* and/or *Z. ardabilae*, suggesting that these introns were recently gained in *Z. tritici*. Hereafter, we will call these the “intron-gain loci.” The other three loci all had the intron presence fixed in *Z. pseudotritici* and *Z. ardabilae*, suggesting that the observed IPAPs in *Z. tritici* are due to recent intron loss (the “intron-loss loci”). We sequenced all eight loci in a global collection of *Z. tritici* and its sister species

doi:10.1093/molbev/msu123 Advance Access publication April 7, 2014



B

	Gene-ID	Characterization	*Z. tritici* Intron	*Z. pseudotritici* Intron	*Z. ardabiliae* Intron
Intron gain	77694	SWIM zinc-finger	IPAP	IPAP	absent
	60105	Glycoside hydrolase family 11	IPAP	absent	absent
	72460	Sugar transporter	IPAP	absent	absent
	72506	Peptidase S8 and S53	IPAP	absent	absent
	84625	Dihydroxy-acid and 6-P-gluconate dehydratase	IPAP	absent	absent
Intron loss	70308	Synaptic vesicle transporter SVOP	IPAP	present	present
	54115	n.a.	IPAP	present	present
	74664	AMP-dependent synthetase and ligase	IPAP	present	present
House-keeping	105948	Actin	present	present	present
	103221	3-isopropylmalate dehydrogenase (leucinB)	present	present	present
	92705	Translation elongation factor EF-1, alpha subunit	present	present	present

Fig. 1. (*A*) Phylogenetic relationships among *Zymoseptoria tritici* and its closest relatives *Z. pseudotritici* and *Z. ardabiliae* collected from undomesticated grasses in Iran. Coalescent times (years before present) inferred in earlier studies are given on the right. (*B*) The nature of the intron polymorphism of the eight assessed genes is listed for every species; fixed presence, fixed absence, and IPAP. Gene-IDs refer to the identification numbers given by the Joint Genome Institute for the fully annotated *Z. tritici* genome of the reference isolate IPO323.

to conduct a population genetic analysis aimed at assessing the relative importance of selection and genetic drift in explaining the recent gains and losses of the corresponding introns in *Z. tritici*.

Results and Discussion

Hypotheses Tested

Following the studies of Llopart et al. (2002) and Omilian et al. (2008), we tested three predictions associated with the hypothesis that natural selection and not genetic drift or demographic history is responsible for the observed IPAPs in *Z. tritici* populations:

1) *Skewed frequency distribution*. The frequency distribution of alleles under selection will be skewed toward rare variants (Braverman et al. 1995). For the five loci that gained introns, we expect to find a skewed distribution only among the intron-present alleles. For the three loci that lost introns, we expect to find a skewed distribution only in the intron-absent alleles.
2) *Reduction in nucleotide diversity*. Beneficial alleles under selection rapidly increase in population frequency. Closely linked sequences are carried along through hitchhiking, leading to a reduction in genetic diversity called a selective sweep across a larger genomic region. New mutations and recombination eventually restore diversity in this region, but these appear slowly, creating an excess of rare alleles compared with other genomic regions (see prediction 1).
3) *Linkage disequilibrium*. Alleles under selection that sweep rapidly through a population will show lower values of recombination rate if there was insufficient time for recombination to decay the resulting linkage disequilibrium (e.g., Sabeti et al. 2002).

Frequencies of IPAPs

Our population genetic analyses identified IPAPs at different stages of fixation. Regional and global frequencies of IPAPs for the eight characterized genes in *Z. tritici* are summarized in figure 2. The relative position of each gene on its respective

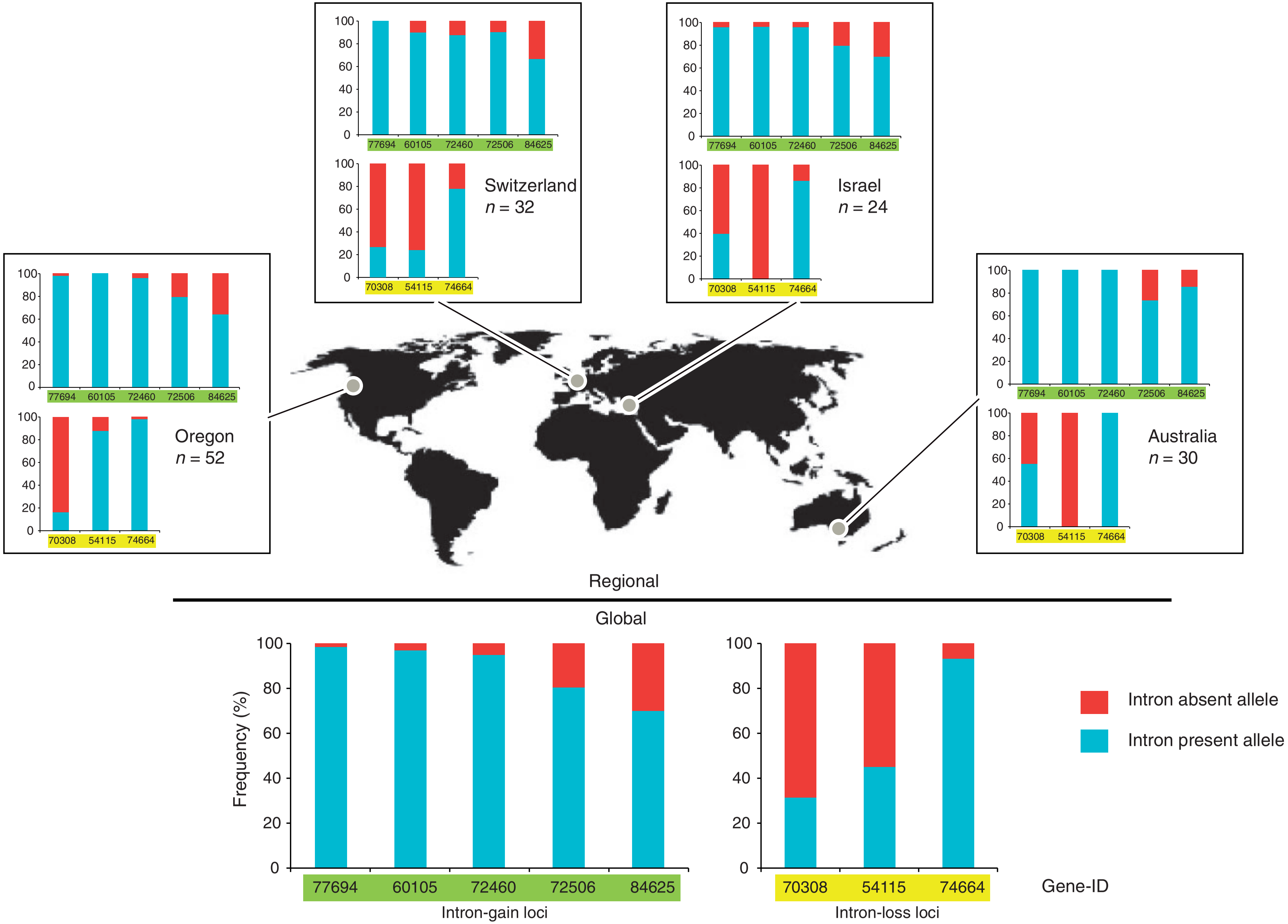


Fig. 2. Sampling locations of *Z. tritici* and frequency distributions of IPAP among the eight assessed genes. Intron gain refers to loci that show a new intron insertion in *Z. tritici* compared with its close ancestors. Intron loss refers to loci with a fixed intron presence among the ancestors but a presence–absence polymorphism in *Z. tritici* (see also fig. 1).

chromosome and the length of exons and introns are given in supplementary figures S2–S10, Supplementary Material online. On the global scale, the intron frequency for the five intron-gain loci ranged from 98% at locus ID-77694 to 70% at locus ID-84625. The intron frequency observed at the three intron-loss loci ranged from 93% at locus ID-74664 to 31% at locus ID-70308. On the regional scale, the frequency of IPAPs was quite variable among *Z. tritici* populations, likely reflecting differences in regional selection pressure and/or differences in demographic histories. Most loci showed IPAPs in all populations, suggesting that drift is probably not a dominant force in driving IPAP to fixation in this species. But in Australia, five of the eight loci were fixed for the presence (IDs-60105, -72460, -77694, -74664) or absence of the respective intron (ID-54115). This observation of reduced intron diversity is consistent with earlier studies that concluded that *Z. tritici* populations in Australia experienced a genetic bottleneck as a result of a recent founder event (Banke et al. 2004; Jürgens et al. 2006).

Intron Sequence Analyses

Mean Tajima's *D* values for the intron sequences were slightly negative but none of them were significantly negative (table 1). Estimated values ranged from $D = +1.24$ to $D = -1.47$, resulting in an overall mean across the eight loci of $D = -0.53$. Similarly, none of the other frequency-based tests for neutrality resulted in significantly negative values for the intron sequences. How do these intron estimates compare with estimates of *D* for the flanking exon regions? Mean Tajima's *D* for the synonymous sites was $D = -0.44$, which is not significantly different from the *D* obtained for the introns ($P = 0.599$, Wilcoxon two-sample test). As nucleotide evolution at synonymous sites is generally regarded as neutral, this suggests an overall neutral evolution for the intron sequences. Estimates of nucleotide diversity π ($\times 10^3$) for introns at the intron-gain loci ranged from 0 to 40.20 with an average of 10.22, which is half the average value estimated for introns at the intron-loss loci (22.19). One hypothesis to explain this difference is that the intron gains were more recent (≤11,000 years before present [ybp]; cf. fig. 1) and had less time to accumulate mutations than the older introns present in the intron-loss loci. Another hypothesis is that the introns and/or the corresponding exons flanking the introns were subjected to stronger directional selection at the intron-gain loci. The latter hypothesis is supported by the introns at loci ID-72460 ($\pi = 0$) and ID-77694 ($\pi = 0.27$) that had the lowest nucleotide diversity (fig. 5), consistent with a rapid sweep hypothesized for the same two loci based on Tajima's *D* values. Because these results suggest that selection did not act on the intron itself, the next section focuses on the analyses conducted on the flanking exon regions.

Table 1. Neutrality Tests for Intron-Gain and Intron-Loss Loci Were Conducted Separately for the Intron-Present (+) and Intron-Absent (−) Alleles.

	T-D	FL-F	FW-Hn
Intron-gain loci IDs			
77 694 +	−2.610***	−4.641***	−5.328***
77 694−	−0.824	0.000	−1.445
60 105 +	0.048	0.116	0.002
60 105−	−0.972	0.959	−1.053
72 460 +	−2.423**	−4.938***	−2.032*
72 460−	2.194*	2.148**	0.129
72 506 +	0.164	0.623	−5.861***
72 506−	1.575	2.006**	−3.646***
84 625 +	−0.108	−0.862	−0.186
84 625−	−0.465	−0.534	−0.176
Intron-loss loci IDs			
70 308 +	−0.587	−1.191	−1.318
70 308−	−0.263	−0.391	−3.735***
54 115 +	−0.758	−0.301	−3.059***
54 115−	1.707	1.590**	−1.384*
74 664 +	−0.173	−2.861**	0.305
74 664−	−0.066	−0.785	0.169

NOTE.—Tests included Tajima's *D* (T-D), Fu and Li's F (FL-F), and the normalized Fay and Wu's *H* (FW-Hn). *, significant deviation from zero at $P < 0.05$; **, significance at $P < 0.01$; ***, significance at $P < 0.001$.

Skewed Frequency Distribution of Rare Alleles

We conducted a number of tests to detect signatures of nonneutral evolution among the intron-gain and intron-loss loci (table 1 and fig. 3). To test our hypothesis of differential selection between intron-present and intron-absent alleles, we conducted separate analyses for these two data partitions for each locus. Analyses of the five intron-gain loci yielded the following results. The average *D* value for the intron-present alleles was moderately negative (−0.91). In contrast, the average value for the intron-absent alleles was moderately positive ($D = 0.74$). This difference, although not significant ($P = 0.174$), is consistent with the hypothesis that selection favored the alleles possessing the intron at the intron-gain loci. More striking is that two of the intron-present alleles showed highly significant negative values (fig. 3*A*) : ID-77694 ($D = -2.60$, $P < 0.001$) and ID-72460 ($D = -2.55$, $P < 0.001$). The corresponding values for the intron-absent alleles of these two loci were moderately positive for ID-77694 ($D = 0.53$) and significantly positive for ID-72460 ($D = 2.14$, $P < 0.01$). Figure 3*B* shows the distribution of Tajima's *D* values across all ~10,000 genes present in the genome of *Z. tritici*. It is clear from this figure that the intron-gain alleles for the intron-gain loci ID-77694 and ID-72460 are located at the extreme end of the distribution, consistent with very strong selection operating on these two loci.

The opposing patterns of distribution of mutations among the intron-present and intron-absent alleles combined with the excess of rare mutations among the intron-present alleles indicate that loci ID-77694 and ID-72460 swept through the *Z. tritici* populations faster than would be expected by drift. Because both of these loci recently gained their introns, these findings indicate that selection strongly favored the intron-present alleles. Interestingly, the coding sequences of the same loci had significantly negative *D* values for both their synonymous and nonsynonymous sites, a pattern consistent with a recent and strong selective sweep affecting the entire locus (supplementary fig. S1, Supplementary Material online).

Frequency-based tests for neutrality can be sensitive to demographic processes. This is particularly true for Tajima's *D* test, where negative values may result from population expansion. We can rule out the possibility that such demographic processes affected our results for several reasons. First, we did not observe an excess of rare mutations (i.e., significantly negative *D* values) in both the intron-absent and the intron-present alleles for the same locus. Second, we obtained significant values for the same loci when Tajima's *D* test was applied to data sets including or excluding the Australian population that was likely affected by demographic events (Banke et al. 2004; Jürgens et al. 2006). Finally, other estimators of neutrality, in particular Fay and Wu's *H* statistic that is more robust to demographic changes because it compares the distribution of high-frequency alleles with intermediate-frequency alleles (Fay and Wu 2000), provided the same outcome.

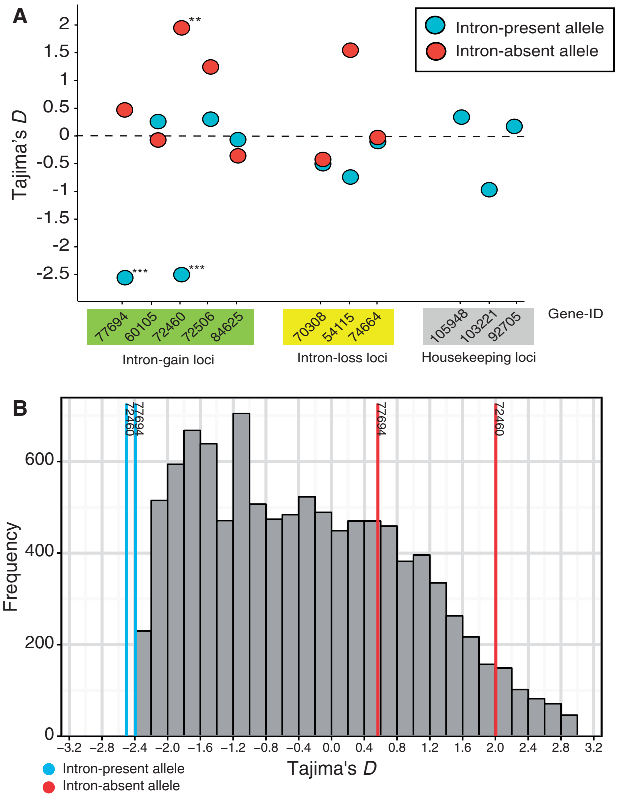


Fig. 3. (*A*) Inferred Tajima's *D* values for the eight assessed loci and different data partitions. Data from each locus were partitioned into intron sequences, exon sequences from intron-present alleles, and exon sequences from intron-absent alleles. *D* values less than zero indicate an excess of low-frequency mutations compared with expectations under neutrality. A significantly negative value of Tajima's *D* is usually interpreted as evidence for rapid population growth or natural selection sweeping beneficial alleles to fixation in a population. **, significant deviation from zero at $P < 0.01$; ***, significant deviation from zero at $P < 0.001$. (*B*) Genome-wide estimates of Tajima's *D* values in *Z. tritici*. The superimposed *D* values for loci IDs-77694 and -72460 show their position at the extreme ends of the distribution.

Assuming a selective advantage for the loss of introns, the expectation is to find signatures of selection at the three intron-loss loci. However, results were less conclusive for these intron-loss loci as none showed significant *D* values. The average value for the intron-absent alleles was slightly positive ($D = 0.59$), whereas the average for the intron-present alleles was slightly negative ($D = -0.68$), suggesting that natural selection did not act on any of the intron-loss loci. However, FW-H estimates were significantly negative for the intron-absent alleles of locus ID-70308 Significant FW-H estimates were also found at locus ID-54115 in both intron-absent and intron-present alleles.

Reduction in Nucleotide Diversity

Our expectation is that the allele classes under putative selection (i.e., the intron-present alleles for the intron-gain loci and the intron-absent alleles for the intron-loss loci) will exhibit lower nucleotide diversity compared with the opposite allele class due to a selective sweep (fig. 5). Consistent with this expectation, estimates of π were lower for intron-present alleles compared with intron-absent alleles at all intron-gain loci (fig. 4; average π $(\times 10^3) = 12.87$ for the intron-present alleles compared with π $(\times 10^3) = 18.08$ for the intron-absent alleles). These differences were significant at three loci that showed signatures of selection in the previous analyses, most notably at locus ID-72460 ($\pi = 2.77$ vs. 38.84). This general pattern of reduction in nucleotide diversity among the intron-gain loci affected both synonymous and nonsynonymous nucleotide positions. Consequently, these selective sweeps resulted in some dramatic reductions in protein diversity, exemplified in figure 5 for locus ID-84625.

In concordance with our finding that none of the intron-loss loci were under selection according to Tajima's *D*, we did not find a generally reduced nucleotide diversity among the intron-absent alleles, with an average $\pi = 40.66$ compared with an average $\pi = 27.27$ among the intron-present alleles. Only locus ID-70308 showed a lower estimate for the intron-absent alleles (14.28 vs. 19.13), but this difference was not significant (fig. 4).

We performed chromosome-wide scans among nine resequenced strains of the Swiss population to identify patterns of decreased nucleotide diversity and/or changes in Tajima's *D* estimator. We found that the intron-gain locus ID-72460 matches exactly a local depression in nucleotide diversity and Tajima's *D*, suggesting a selective sweep originating from this locus and a larger depression in nucleotide diversity within 50 kb (supplementary fig. S4, Supplementary Material online). Three other intron-gain loci showed depressed nucleotide diversity and low Tajima's *D* within 50–70 kb of the locus (ID-60105, 72506, and 84625). The intron loss locus ID-70308 showed a similar pattern of depressed nucleotide diversity and Tajima's *D* within 50 kb (supplementary fig. S7, Supplementary Material online).

Linkage Disequilibrium

A rapid selective sweep will result in linkage disequilibrium at the selected locus, which will result in lower rates of recombination compared with loci not under directional selection (Barton 2000). The average estimated recombination rate for the intron-present alleles among the intron-gain loci was 20 times lower than found for the intron-absent alleles (3.69×10^{-4} vs. 7.46×10^{-3}). Consistent with the previous analyses of nucleotide diversity, recombination rates for intron-present alleles at all loci were significantly lower compared with the intron-absent alleles (fig. 6). In contrast, we found very similar recombination rates at the three intron-loss loci for both intron-absent and intron-present alleles, consistent with an absence of selection at these loci.

Comparison to Housekeeping Genes

We conducted the same analyses applied to the IPAP loci on global data sets of three housekeeping genes, that is, genes encoding actin, elongation factor, and leucine biosynthesis (fig. 1). As these loci are not likely to be affected by recent selective processes, they provide a neutral background for comparison against the IPAP loci.

All three housekeeping loci showed Tajima's *D* values not significantly different from zero, consistent with our assumption of neutrality (fig. 3). Average nucleotide diversity for the intron sequences was π ($\times 10^3$) = 40.33, 2× higher than the average for the intron-loss loci and 4× higher than the average for the intron-gain loci (fig. 4). This pattern of higher nucleotide diversity at the intron sites of the housekeeping genes likely reflects the long-term accumulation of neutral mutations and the absence of recent selective sweeps affecting these housekeeping genes. The pattern also is consistent with recent selective sweeps affecting the intron-gain loci and to a lesser extent the intron-loss loci. In contrast, nucleotide diversity for the exon sequences of the housekeeping genes was lower than the exon diversity observed for the intron-gain loci (fig. 4; average π ($\times 10^3$) = 7.09). This finding of low exon diversity and high intron diversity is expected for genes coding for proteins that are under strong functional constraints, such as housekeeping genes.

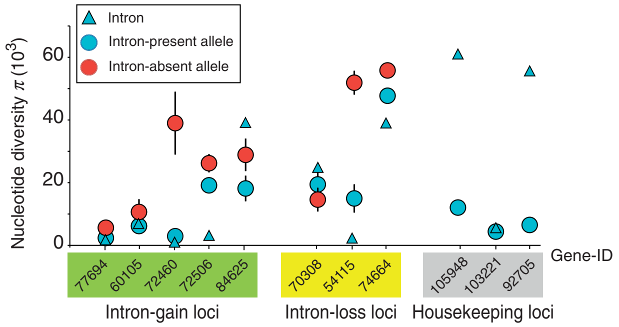


Fig. 4. Estimates of nucleotide diversity (π) for the eight assessed loci and different data partitions. Data from each locus were partitioned into intron sequences, exon sequences from intron-present alleles, and exon sequences from intron-absent alleles. Error bars represent two times the standard deviation.

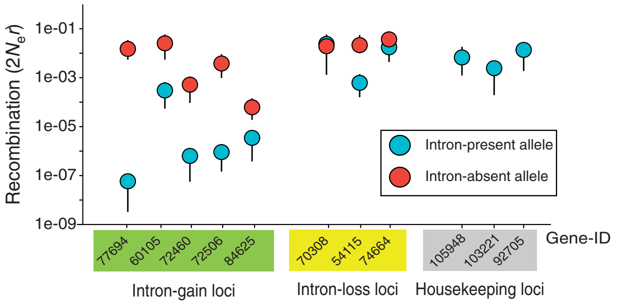


Fig. 6. Coalescent-based estimates of recombination rates. The overall recombination rate *r* is the product of the recombination rate per site per generation and the neutral mutation rate per site per generation. Bars represent the 95% confidence intervals.

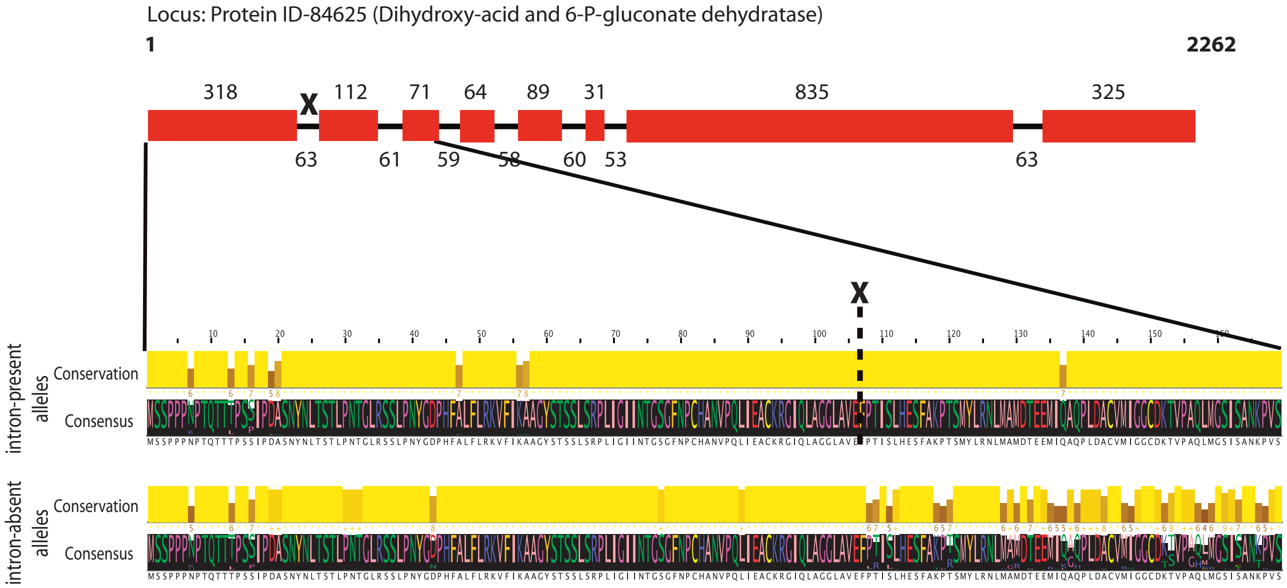

Locus: Protein ID-84625 (Dihydroxy-acid and 6-P-gluconate dehydratase)
1
2262
318
112
71
64
89
31
835
325
X
63
61
59
58
60
53
63
X
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
intron-present alleles
Conservation
Consensus
MSSPPPNPTQTTTPSSIPDASNYLNTSTLPNTGLRSSLPNYGDPHFALFLRKVFIKAAGYSTSSLSRPLIGIINTGSGFNPCHANVPQLIEACKRGIQLAGGLAVEFPTISLHESFAKPTSMYLRNLMAMDTEEMIQAQPLDACVMIGGCDKTVPAQLMGSISANKPVS
intron-absent alleles
Conservation
Consensus
MSSPPPNPTQTTTPSSIPDASNYLNTSTLPNTGLRSSLPNYGDPHFALFLRKVFIKAAGYSTSSLSRPLIGIINTGSGFNPCHANVPQLIEACKRGIQLAGGLAVEFPTISLHESFAKPTSMYLRNLMAMDTEEMIQAQPLDACVMIGGCDKTVPAQLMGSISANKPVS

Fig. 5. Schematic representation of locus ID-84625 showing the dramatic reduction in amino acid diversity in intron-present alleles compared with intron-absent alleles. Crosses mark the position of IPAP.

As expected, analyses of linkage disequilibrium did not indicate recent selective sweeps occurring at the housekeeping loci. Accordingly, estimates of recombination rates were in the range of the highest values observed at the IPAP loci (fig. 6). The average estimate of 1.07×10^{-2} was very similar to the average estimate across the intron-loss loci (1.29×10^{-2}) and the intron-absent alleles of the intron-gain loci (7.46×10^{-3}), supporting our conclusion that these alleles were not affected by a selective sweep.

Transcription Analyses

Expression results for all genes were obtained from an RNAseq experiment using the Swiss isolate 3D7 and are summarized in supplementary figure S10 (Supplementary Material online). Consistent with the conserved functional properties associated with housekeeping genes, all three genes were expressed constitutively at approximately the same level across the different life-cycle stages of *Z. tritici*. In contrast, most of the IPAP loci showed significant differences in expression during different stages in the life cycle. For example, the intron-gain locus ID-72460 that shows signatures of a selective sweep was expressed ~3× more in the necrotrophic phase compared with the biotrophic phase and was expressed ~10× less in the saprotrophic phase compared with the biotrophic phase. Similar patterns of differential expression were reported previously for several cell-wall degrading enzymes in *Z. tritici*, and associated with selection for host adaptation and specialized functions during different life-cycle stages (Brunner et al. 2013). We did not find evidence for alternative splicing and resulting isoforms of the corresponding proteins for any of the IPAP loci. Our RNAseq experiment however only includes one isolate, and we cannot rule out that other isoforms are transcribed in other isolates.

Alternative Evolutionary Scenarios

So far, we made the parsimonious assumption that introns present in *Z. tritici* but absent in the two sister species *Z. pseudotritici* and *Z. ardabiliae* represent intron gains in *Z. tritici* (fig. 1). However, an alternative evolutionary scenario is that the intron-gain loci actually represent ancestral intron-presence alleles, where the introns were lost independently in the sister species. We consider this violation of parsimony as very unlikely given the results of our analyses. Our interpretation of a recent intron gain at these loci implies a reduced accumulation of nucleotide diversity at the intron sites compared with the respective exon regions due to their shorter evolutionary history. The generally lower nucleotide diversity of the introns compared with the respective exon regions is shown in figure 4. For example, the nucleotide diversity of the exon regions of loci ID-72460 and ID-77694 are 0.0027 and 0.0017. The corresponding intron diversities are clearly lower with values of 0.0000 and 0.0002, respectively.

To further test the alternative hypothesis, we conducted coalescent-based analyses to specifically test the intron-gain hypothesis by simulating times to the most recent common ancestors (TMRCA) for introns and exons separately. Because there was no nucleotide variation for the intron in gene ID-72460, we could only compare gene ID-77694 with one of the housekeeping genes. The results show nonoverlapping distributions of TMRCA, indicating significantly different coalescent times for intron and exon sequences. Furthermore, the estimates of TMRCA for the intron sequences are much younger (i.e., closer to zero/more recent), than the estimates for the exon sequences, thus fully supporting our assumption of intron gain at this locus. In contrast and as expected, the housekeeping gene largely shows overlapping distributions of TMRCA for introns and exons, suggesting a long common coalescent history for these two gene regions (supplementary fig. S11, Supplementary Material online).

Conclusions

We conducted population genetic analyses to assess the contribution of selection to the within-species IPAPs found in natural populations of the plant pathogenic fungus *Z. tritici* and its two sister species. None of the intron-loss loci appeared to be under selection. In contrast, we found compelling evidence that at least two out of the five intron-gain loci, ID-77694 (SWIM zinc-finger) and ID-72460 (sugar transporter) are under directional selection. These findings suggest that the intron-present alleles of these loci are sweeping to fixation in the global population, providing a hitchhiking mechanism to explain the rapid intron gain observed in *Z. tritici*.

Materials and Methods

In a recent study, we utilized population genomic comparisons to identify transitory phases of intron gains and losses in three closely related fungi (Torriani et al. 2011). We found 74 intron positions showing intraspecific presence–absence polymorphisms for the entire intron. For this study, we selected eight of these loci showing IPAPs in *Z. tritici* (fig. 1). We sequenced five genes that exhibited a rapid intron gain in *Z. tritici* compared with its ancestor *Z. ardabiliae* (Gene IDs-60105, -72460, -72506, -77694, and -84625) and three genes that showed a rapid intron loss (Gene IDs-54115, -70308, and -74664). Gene IDs were assigned according to the annotated *Z. tritici* reference genome isolate IPO323 (Goodwin et al. 2011).

Zymoseptoria Isolates, DNA Isolation, and Sequencing

All *Zymoseptoria* isolates used in this study originated from natural infections (fig. 2). The *Z. tritici* isolates were collected from four geographically distant wheat fields: one each from Oregon, USA ($n = 53$), Australia ($n = 30$), Switzerland ($n = 32$), and Israel ($n = 24$). These field populations were characterized previously using restriction fragment length polymorphisms and quantitative traits (Zhan et al. 2005). The isolates from *Z. pseudotritici* ($n = 26$) and *Z. ardabiliae* ($n = 33$) were collected in Iran on wild grasses (Stukenbrock et al. 2007). As a comparison for intron evolution at loci showing no IPAP, we analyzed sequences from three previously described housekeeping genes, that is, genes encoding actin, elongation factor, and leucine biosynthesis (Banke and McDonald 2005; Stukenbrock et al. 2007; Stukenbrock et al. 2012).

Primers used to polymerase chain reaction (PCR)-amplify the genes were designed to include the exon sequences flanking the intron site in both directions. PCR amplicons were sequenced in both directions using the same primers on an ABI 3730 xl sequencer (Applied Biosystems). Obtained nucleotide sequences were aligned using MAFFT (Multiple Alignment using Fast Fourier Transform; Katoh et al. 2009) and edited and partitioned using BIOEDIT (Hall 1999) to generate three data sets: intron sequences, intron-present alleles, and intron-absent alleles.

We obtained whole-genome resequencing single nucleotide polymorphism data for nine resequenced isolates of the Swiss population in order to perform chromosome-wide scans of nucleotide diversity (Croll et al. 2013).

Population Genetic Analyses

We used the program DnaSP v5 (Librado and Rozas 2009) to calculate population genetic parameters and estimate deviation from neutral expectations for the three data partitions. Nucleotide diversity was estimated as π, the average number of nucleotide differences per site between two sequences (Tajima 1983). Neutrality tests included four frequency-based tests of allelic variation: Tajima's *D* (Tajima 1989), Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000). The tests of Fu and Li and Fay and Wu were conducted using the closely related sister species *Z. ardabiliae* and *Z. pseudotritici* as outgroups.

Recombination rates between alleles containing or lacking the intron were estimated using the full-likelihood and coalescence-based method implemented in the program LAMARC v2.1.8 (Kuhner 2006). Parameters estimated by LAMARC included θ, the effective number of individuals, and *r*, the product of the recombination rate per site per generation and the neutral mutation rate per site per generation. From these parameters, we obtained the population recombination rates $2N_er = \theta \times r$. Felsenstein 84 (F84) was used as the nucleotide substitution model, and the transition/transversion ratios were set individually for each gene according to estimates obtained using the program MEGA (Tamura et al. 2011).

Expression Analyses

The hemibiotrophic pathogen *Z. tritici* displays a life cycle composed of three phases: an initial biotrophic and asymptomatic phase spanning 10–12 days post-inoculation (dpi), followed by a necrotrophic period lasting 1–4 days during which the wheat leaf tissue is killed, and a saprotrophic phase during which the pathogen finishes its reproductive cycle and lives as a saprophyte on the dead leaves. We recently conducted RNAseq analyses (Morin et al. 2008) to determine genome-wide gene expression profiles during each phase in the pathogen life cycle. A detailed description of the experiment is given in Brunner et al. (2013). From this experiment, data for the eight IPAP loci and the three housekeeping genes were extracted to confirm their expression in planta and to assess their relative levels of expression during the pathogen life cycle.

Coalescent Analyses

We conducted coalescent-based analyses to specifically test the intron-gain hypothesis by simulating TMRCA for introns and exons separately. The expectation was that recently gained introns will show a more recent coalescent time than the corresponding exon sequences. The TMRCA of the intron and exon sequences was inferred using the coalescent-based Bayesian Markov chain Monte Carlo (MCMC) method implemented in the program BEAST version 1.4.1 (Drummond and Rambaut 2007). A strict molecular clock model was applied with "exponential growth" as the tree prior. Using other priors (e.g., relaxed molecular clock) did not change the overall results. To allow a direct comparison with the neutral intron evolution, only variation at silent sites from the exon regions were included in the analyses. The MCMC analyses were run for 10^7 generations, sampling every 1,000th iteration after an initial burn in of 10%. The performance of the MCMC process was checked for stationarity and large effective sample sizes in TRACER (available from http://beast.bio.ed.ac.uk/software/tracer/, last accessed April 18, 2014). The 95% credibility intervals of the estimated TMRCAs were depicted using TRACER.

Supplementary Material

Supplementary figures S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are particularly grateful to Renato Guidon for collecting the sequence data. The Genetic Diversity Center at ETH Zurich (GDC) provided sequencing facilities. This work was supported by ETH Zurich.

References

Banke S, McDonald BA. 2005. Migration patterns among global populations of the pathogenic fungus *Mycosphaerella graminicola*. *Mol Ecol.* 14:1881–1896.

Banke S, Peschon A, McDonald BA. 2004. Phylogenetic analysis of globally distributed *Mycosphaerella graminicola* based on three DNA sequence loci. *Fungal Genet Biol.* 41:226–238.

Barton NH. 2000. Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci.* 355:1553–1562.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140:783–796.

Brunner PC, Torriani SFF, Croll D, Stukenbrock EH, McDonald BA. 2013. Coevolution and life cycle specialization of plant cell wall degrading enzymes in a hemibiotrophic pathogen. *Mol Biol Evol.* 30:1337–1347.

Croll D, Zala M, McDonald BA. 2013. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet.* 9:e1003567.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Eyal Z. 1999. The *Septoria tritici* and *Stagonospora nodorum* blotch diseases of wheat. *Eur J Plant Pathol.* 105:629–641.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Fu Y-X, Li W-H. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Goodwin SB, Ben M'Barek S, Dhillon B, Wittenber AHJ, Crane CF, Hane JK, Foster AJ, Van der Lee TAJ, Grimwood J, Aerts A, et al. 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* 7(6):e1002070.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.* 41:95–98.

Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet.* 22:16–22.

Jürgens T, Linde CC, McDonald BA. 2006. Genetic structure of *Mycosphaerella graminicola* populations from Iran, Argentina and Australia. *Eur J Plant Pathol.* 115:223–233.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol.* 537:39–64.

Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22:768–770.

Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in *Daphnia* populations. *Science* 326:1260–1262.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad Sci U S A.* 99: 8121–8126.

Lynch M. 2002. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A.* 99:6118–6123.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, Marra MA. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45:81–94.

Omilian AR, Scofield DG, Lynch M. 2008. Intron presence-absence polymorphisms in *Daphnia*. *Mol Biol Evol.* 25:2129–2139.

Orton ES, Deller S, Brown JKM. 2011. *Mycosphaerella graminicola:* from genomics to disease control. *Mol Plant Pathol.* 12:413–424.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet.* 7:211–221.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Stukenbrock EH, Banke S, Javan-Nikkha M, McDonald BA. 2007. Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Mol Biol Evol.* 24:398–411.

Stukenbrock EH, Quaedvlieg W, Javan-Nikhah M, Zala M, Crous PW, McDonald BA. 2012. *Zymoseptoria ardabiliae* and *Z. pseudotritici*, two progenitor species of the septoria tritici leaf blotch fungus *Z. tritici* (synonym: *Mycosphaerella graminicola*). *Mycologia* 104: 1397–1407.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Torriani SFF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D. 2011. Evidence for extensive recent intron transposition in closely related fungi. *Curr Biol.* 21:2017–2022.

Zhan J, Linde CC, Juergens T, Merz U, Steinebrunner F, McDonald BA. 2005. Variation for neutral markers is correlated with variation for quantitative traits in the pathogenic fungus *Mycosphaerella graminicola*. *Mol Ecol.* 14:2683–2693.

Zhan J, Pettway RE, McDonald BA. 2003. The global genetic structure of the wheat pathogen *Mycosphaerella graminicola* is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. *Fungal Genet Biol.* 38:286–297.