Title: The genomic rate of adaptation in the fungal wheat pathogen *Zymoseptoria tritici*

Authors: Jonathan Grandaubert\(^1,2\), Julien Y. Dutheil\(^3,4,5\), Eva H. Stukenbrock\(^1,2,5\)

Affiliations:
\(^1\) Environmental Genomics Group, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany, \(^2\) Christian-Albrechts University of Kiel, Am Botanischen Garten 1-9, 24118 Kiel, Germany, \(^3\) Research group Molecular Systems Evolution, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany, \(^4\) ISEM – Institut des Sciences de l’Évolution. Phylogeny and Molecular Evolution. Place Eugène Bataillon Cc 065, 34095 Montpellier cedex 05, France

\(^5\) Corresponding authors:
Eva H. Stukenbrock, Environmental Genomics, CAU Kiel, Am Botanischen Garten 1-11, 24118 Kiel, Germany, Email: estukenbrock@bot.uni-kiel.de, phone: +49 (0) 431 880 6368
Julien Y. Dutheil, Research group Molecular Systems Evolution, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany, Email: dutheil@evolbio.mpg.de, phone: +49 (0) 4522 763 298

Keywords: Genome evolution, adaptation, evolutionary rates, recombination, plant pathogenic fungi
Abstract:
Antagonistic host-pathogen co-evolution is a determining factor in the outcome of infection and shapes genetic diversity at the population level of both partners. Little is known about the overall genomic rate of evolution in pathogens or the molecular bases of rapid adaptation. Here we apply a population genomic approach to infer genome-wide patterns of selection among thirteen isolates of the fungal pathogen Zymoseptoria tritici. Based on whole genome alignments, we report extensive presence-absence polymorphisms of genes resulting from a high extent of within population karyotypic variation. We apply different test statistics based on the distribution of non-synonymous and synonymous polymorphisms (pN/pS) and substitutions (dN/dS) to estimate rates of adaptation and identify specific targets of selection. Overall we estimate that 44% of substitutions in the proteome of Z. tritici are adaptive and we find that the rate of adaptation is positively correlated with recombination rate. The proportion of adaptive amino acid substitutions in genes encoding determinants of pathogenicity is as high as 68% underlining the importance of positive selection in the evolution of virulence-associated traits. Using a maximum likelihood approach and codon models we furthermore identify a set of 786 additional genes showing signatures of diversifying selection. We furthermore addressed the effect of different parameters on adaptation using a linear model with pN/pS as response variable. Our results point to an increased effectiveness of purifying selection with increased recombination rate and high gene density. Overall we show a central role of sexual recombination in protein evolution in this important wheat pathogen.
**Introduction**

Antagonistic host-pathogen interactions drive co-evolutionary dynamics between pathogens and their hosts. Signatures of selection in genome data can inform about mechanisms of evolution and identify targets of selection at interacting loci (Möller and Stukenbrock 2017). In general, genome studies of microbial pathogens have focused on rapidly evolving genes involved in pathogenicity, such as “effector” genes encoding proteins that interfere with host defences and may determine host range of the pathogen (Lo Presti et al. 2015). Effector genes of fungal pathogens have frequently been found to associate with repetitive DNA and it is proposed that repeat rich genome compartments provide particularly favourable environments for rapid evolution of new virulence specificities (e.g. (Ma et al. 2010; Spanu et al. 2010; Klosterman et al. 2011; Daverdin et al. 2012)). Repetitive DNA may locally increase mutation rate and contribute to gene duplications and structural variation among alleles (Dutheil et al. 2016). Yet little is known about the interplay of mutation, natural selection, genetic drift and, for sexually reproducing species, recombination along the genome of fast evolving pathogens.

Rates of evolution and the type of selection acting on a gene can be inferred from the comparison of sequences sampled from distinct individuals in a population. Using statistical population genetics, it is possible to infer the rate at which certain mutations are fixed in comparison to the ‘null’ expectation of neutral evolution, that is, without selection (Nei and Kumar 2000). In the case of protein-coding sequences, this neutral rate is typically given by the rate of synonymous substitutions, hereby assumed to be selectively neutral. The relative proportion of synonymous and non-synonymous substitutions thus provides a signature of the underlying type of selection acting on a gene. An excess of segregating or fixed non-synonymous mutations is indicative of positive selection while an absence or reduced proportion of non-synonymous mutations shows conservation of the sequence at the amino acid level indicative of purifying selection (Nei and Gojobori 1986). Further tests, building on pioneering work by McDonald and Kreitman (McDonald and Kreitman 1991), assess adaptive evolution by comparing the ratio of non-synonymous to synonymous substitutions between species, dN/dS, to the ratio of non-synonymous to synonymous polymorphisms within species, pN/pS. In the absence of selection, the dN/dS ratio is expected to be equal to the pN/pS ratio. However, for a gene that has undergone positive selection, the dN/dS ratio is expected to be higher than the pN/pS ratio. This is because rare, positively selected mutations represent only a minor proportion of all polymorphisms in a population sample. However, positively selected mutations have a higher probability of fixation than neutral
mutations and will therefore accumulate over time as species diverge, resulting in a higher dN/dS ratio (Eyre-Walker 2006; Galtier 2016).

The McDonald-Kreitman table forms the basis for the quantification of the proportion of adaptive substitutions in a gene fixed by positive selection defined as $\alpha = 1 - (pN / pS) / (dN / dS) = 1 - (dS * pN) / (dN * pS)$ (Smith and Eyre-Walker 2002). Additional modeling of the fitness effect of substitutions allows $\alpha$ to account for the presence of slightly deleterious or slightly advantageous mutations that otherwise can bias the inference of positively selected substitutions (Eyre-Walker and Keightley 2007). These methods have been applied to a range of different species and have been used to show that the rate of adaptive evolution varies largely among different taxa (Eyre-Walker 2006; Galtier 2016). Overall, the variation in $\alpha$ correlates positively with the average effective population size (Ne), where species with large Ne more efficiently remove slightly deleterious mutations resulting in a reduced pN value and lower pN/pS ratios compared to species with small Ne (Galtier 2016).

In pathogen genomes, increased rates of adaptation, as occurring by arms-race evolution where new non-synonymous mutations repeatedly are selected for, will result in higher dN/dS ratios. Indeed, evolutionary predictions based on dN/dS analyses and McDonald-Kreitman tests have identified effectors and other pathogenicity-related genes in pathogen genomes (Raffaele et al. 2010; Stukenbrock et al. 2011). However, genes that evolve by trench warfare evolution leave different signatures in the genome sequence. This type of selection can also be referred to as positive diversifying selection where a diversity of alleles at the selected locus is maintained by natural selection. The genic signature of positive diversifying selection is an increased proportion of non-synonymous polymorphisms (Charlesworth 2006). Yet, in a functionally important gene, the majority of non-synonymous sites will be evolving under selective constraints, and only a small number will be allowed to change without negatively affecting protein function. Therefore, in most cases, searching for signatures of positive diversifying selection simply by comparing the average proportion of non-synonymous to synonymous polymorphisms in a gene will show little statistical power (Nielsen and Yang 1998). To circumvent this, evolutionary codon models that allow for varying selection intensity across a gene sequences can be applied to identify specific codon positions that have an excess of non-synonymous to synonymous mutations (a higher pN/pS ratio) (Nielsen and Yang 1998).

In this study we assess the different types of selection affecting the genome of a fungal wheat pathogen *Zymoseptoria tritici* (synonym *Mycosphaerella graminicola*). The genome of this pathogen consists of thirteen core chromosomes and several accessory chromosomes that
are present at different frequencies in different individuals, constituting a particular case of
caryotypic polymorphism (Goodwin et al. 2011). The accessory chromosomes comprise
repeat rich, heterochromatic DNA with a low gene content and a so far unknown functional
relevance (Grandaubert et al. 2015; Schotanus et al. 2015). We have previously shown that
evolutionary rates on these chromosomes are particularly high suggesting that genes on the
accessory chromosomes in general evolve under less selective constraints (Stukenbrock et al.
2010).

Previous studies based on comparative population genomic analyses of Z. tritici and two
closely related species, Z. pseudotritici and Z. ardabiliiae used genome-wide estimates of non-
synonymous and synonymous divergence to identify past species-specific signatures of
selection in the wheat pathogen (Stukenbrock et al. 2010; Stukenbrock et al. 2011).

Functional characterization of some of these genes revealed amino acid substitutions
important for in planta development and asexual spore formation in the wheat-adapted
pathogen, and thereby confirmed the use of evolutionary predictions to identify functionally
relevant traits (Poppe et al. 2015a).

While past signatures of selection can be related to ecological specialization of Zymoseptoria
species, ongoing positive selection likely reflects host-pathogen arms race or trench warfare
evolution and recurrent adaptations to changing environmental conditions including fungicide
treatments and wheat cultivars. Population genomic studies of Z. tritici have revealed that the
pathogen has a high effective population size, which correlates positively with a strong
efficacy of natural selection, i.e. the ability to rapidly fix advantageous mutations and
eliminate non-advantageous (Stukenbrock et al. 2011). Additional parameters that can
influence rapid adaptation are recombination and mutation rates, as well as genome
architecture. Recombination rate variation has been studied in Z. tritici by experimental
mating and population genomic sequencing. Independently, these approaches have
confirmed exceptionally high recombination rates (~60cM/Mb) (Croll et al. 2015;
Stukenbrock and Dutheil 2017)). So far the implications of such high recombination rates on
gene evolution are not known.

We apply a population genomics approach to infer genome-wide positive and purifying
selection among thirteen isolates of Z. tritici collected from bread wheat in Europe and the
Middle East. Our analyses based on more than 1.4 million single nucleotide polymorphisms
(SNPs) and including 700,000 coding sites allow us to identify past and on-going signatures of
selection in the genome of Z. tritici. Notably, our analyses reveal a strong importance of
recombination in gene evolution and a particularly high rate of adaptive substitutions in
genes encoding putative effectors. These analyses underline the potential of rapid adaptation of virulence related traits in this important agricultural pathogen.

Results and Discussion

Genetic structure of Z. tritici populations

We generated a population genomic dataset of thirteen Z. tritici isolates obtained from different field populations in Europe and Iran (Table S1). Given the high extent of structural variation in genomes of Z. tritici isolates, we assembled each genome de novo to capture structural differences. We used the resulting thirteen genome assemblies to generate a filtered multiple genome alignment of ~27 Mb (Table S2) comprising a total of 1,489,362 SNPs of which approximately 50% locate in protein coding regions (Table S1). The SNP data was used to compute the overall genetic diversity of the sample showing a mean value of \( \pi = 0.022 \) per site.

Several studies have addressed the population genetic structure of Z. tritici at different spatial scales (e.g. (Linde et al. 2002; BANKE and MCDONALD 2005)). These studies have concluded that Z. tritici comprises a panmictic population with very little population structure. We here used the genomic data to investigate the relationship of the Z. tritici isolates. First, we assessed population genetic structure by analyzing the ancestral recombination graph of the thirteen genomes. To this end, we slid 10 kb windows along the multiple genome alignment, and estimated the genealogy for each window. The resulting 1,850 trees were combined into a super tree (Figure 1A). If the sample of genomes is taken from a panmictic population, the super tree is expected to be a star tree. However, here we observe at least two clusters, one comprising all European isolates and the other comprising two isolates collected in Iran (Figure 1A). We further investigated population structure using the program ADMIXTURE (Alexander et al. 2009) and found the strongest support for a model with two ancestral populations supporting the tree-based clusters of European and Iranian isolates (Figure 1B). Within the European population four German isolates furthermore form an independent group also clustered together by the ADMIXTURE analysis using k=3. While the population sample used here is smaller than the sample sizes used in previous population genetic studies, our analyses build on a considerably larger number of variable sites providing a higher resolution of the genealogical relationships among Z. tritici isolates. The thirteen isolates show genetic differentiation that reflects their geographical origin suggesting some extent of subdivision and geographical barriers to gene flow between populations. This geographical pattern is consistent with the global genetic structure of other important wheat
pathogens such as the wheat yellow rust fungus *Puccinia striiformis f.sp. tritici* and the wheat scab pathogen *Fusarium graminearum* (O’Donnell et al. 2000; Ali et al. 2014), and shows that local adaptation and population divergence can shape the genetic structure of cereal pathogens even in the presence of global agricultural trade and long distance pathogen dispersal (Brown and Hovmøller 2002).

**Z. tritici** shows presence-absence polymorphism of genes and whole chromosomes

*Z. tritici* exhibits dramatic patterns of chromosome length polymorphisms and presence-absence variation of accessory chromosomes (Mehrabi et al. 2007; Brunner et al. 2013; McDonald et al. 2016). Based on the multiple genome alignment, we here investigated presence-absence polymorphisms of chromosomes, chromosome fragments and genes among the thirteen isolates. We focused on non-repetitive DNA and show, in agreement with a previous study (Croll et al. 2013), that 96.4% of the core chromosomes (chromosome 1 to 13) are shared among all isolates (Figure S1). However, as expected, the mapping of assembled contigs aligning to the accessory chromosomes varied considerably between isolates (Figure S1).

Overall, 9,499 of the 11,839 genes predicted in the reference genome of IPO323 are present in all of the thirteen isolates. Of the remaining, 1,584 genes are located on core chromosomes and 667 on accessory chromosomes. This set of genes comprises genes present at different frequencies among the thirteen *Z. tritici* isolates including 89 genes present only in IPO323 and of which 77 locate on core chromosomes.

A previous genome comparison of two Swiss *Z. tritici* isolates revealed nearly 300 novel genes not characterized in the reference genome (Plissonneau et al. 2016). We speculated that the different isolates of *Z. tritici* also could contain accessory genes and chromosomes not present in the reference IPO323. To test this we investigated the subset of contigs that were not aligned to the reference genome. We found a total of 5.2 Mb in the twelve re-sequenced *Z. tritici* isolates with no homology to sequences in the reference genome. These new *Z. tritici* sequences, corresponding to accessory fragments, encompass between 602 kb and 995 kb of non-repetitive DNA in each genomes. We applied an *ab initio* gene predictor to assess whether these accessory sequences encode proteins and found evidence for 1,041 novel full-length gene models encoding proteins of at least 30 amino acids. Of these 75% encode proteins without known function consistent with a general pattern observed on the accessory chromosomes with a high proportion of gene models without a predicted function [19]. Isolate-specific fragments were previously shown to encode effector candidate genes strongly
up-regulated in *Z. tritici* during host infection (Plissonneau et al. 2016). In other fungal pathogens such as *Verticillium dahliae* and *Magnaporthe oryzae* isolate-specific genes were experimentally shown to encode virulence determinants (de Jonge et al. 2012; Xue et al. 2012). Based on our findings, we hypothesize that the large complement of isolate-specific genes in thirteen *Z. tritici* genomes could play a role in host-pathogen interactions. Under this scenario, the observed variation in gene content could reflect diversifying selection imposed by genetic traits in a variety of host genotypes.

**Measures of adaptive evolution in *Z. tritici***

We next aimed to obtain a quantitative assessment of adaptive mutations in the genome of *Z. tritici*. To this end, we first estimated the non-synonymous and synonymous divergence $dN$ and $dS$ using a genome alignment of *Z. tritici* and its sister species *Z. ardabiliae*. Furthermore, we used the *Z. tritici* SNP data to compute the unfolded site frequency spectra (SFS) of synonymous and non-synonymous sites using *Z. ardabiliae* as an outgroup. By contrasting divergence and polymorphism data, we estimated the parameter $\alpha$ as a measure of the proportion of adaptive non-synonymous substitutions. The SFS is however strongly affected by demography and the presence of slightly deleterious mutations segregating at low frequencies (see review by (Eyre-Walker and Keightley 2007)). State-of-the-art statistical methods account for these effects by modeling the distribution of fitness effects (DFE) of mutations (Gossmann et al. 2010; Galtier 2016). We used the program Dfem to compute $\alpha$ as well as $\omega_n$ the rate of adaptive substitutions, using four distinct DFE models accounting for mutations with both slightly deleterious and beneficial effects (Galtier 2016). We fitted and compared the DFE models (see Materials and Methods) and found that the Gamma-Exponential model fitted our data best (Table 1) in agreement with studies of genetic data from animals (Galtier 2016). This suggests the existence of slightly deleterious, as well as slightly beneficial segregating mutations in the genome of *Z. tritici* (Table 1). These estimates based on a Gamma-Exponential model suggest an $\alpha$ value of 44% as a genome average, and an $\omega_n$ value of 0.05. Both values are in the range of what is observed for Mammals (with the exception of Primates) but considerably higher than estimates from plants (Gossmann et al. 2010; Galtier 2016). In candidate effector genes, however, the rate of mutations fixed by selection is more than twice as high as the average of all genes (Table 1), with 68% of substitutions being estimated as adaptive. This estimate is close to the highest values reported in animals (Galtier 2016). We further aimed at assessing the impact of recombination on the fixation of adaptive mutations, by considering genes in distinct
recombination environments. We report a significant positive correlation between \( \alpha \) and \( \rho \)
(the population scaled recombination rate inferred as \( \rho = 4Ne \times r \)) (Kendall’s tau = 0.25, p-
value = 6.816e-6) and \( \omega_0 \) and \( \rho \) (Kendall’s tau = 0.21, p-value = 1.95e-4, Figure 2), suggesting
that a higher recombination rate favors the fixation of adaptive mutation, as expected by a
Hill-Robertson effect (Hill and Robertson 1966; Marais and Charlesworth 2003). These results,
based on population genomic data of \( Z. \) tritici, represent the first quantitative assessment of
adaptive evolution in the genome of a fungal pathogen. The exceptionally high rate of
adaptation in effector genes likely reflects arms race evolution driven by the antagonistic
interaction of \( Z. \) tritici and its host.

Strength of purifying selection along the genome of \( Z. \) tritici

Loci at different positions of the genome evolve at different speed. This is due to variation in
the effect of different evolutionary processes across the genome as well as variation in the
composition of the DNA sequence itself (Ellegren 2014). In sexual organisms, recombination is
a strong determinant of evolutionary rates, and differences in the frequency of recombination
can strongly impact the rate whereby adaptive mutations can be fixed and whereby (slightly)
deleterious mutations are removed (Marais and Charlesworth 2003). Evolutionary processes
at any given site in the genome are strongly influenced by selection acting on linked sites, a
process referred to as linked selection, including Hill-Robertson interference in the case of
linkage with positively selected mutations (Hill and Robertson 1966) and background
selection in the case of linkage with mutation under purifying selection (Hill and Robertson
1966; Nordborg et al. 1996). Here, we aimed to address the genome-wide strength of
purifying selection in \( Z. \) tritici in protein coding genes, using the ratio of non-synonymous to
synonymous polymorphisms (pN/pS ratio).

We computed pN and pS for each gene with two methods: (1) by inference of average
pairwise heterozygosity (Romiguier et al. 2014; Ellegren and Galtier 2016) and (2) using a
maximum likelihood framework and models of sequence evolution (Nielsen and Yang 1998).
In this framework, two types of models were compared: a two-rate model (M1a), with sites
evolving under neutrality or purifying selection with a pN/pS ratio equal or below one (neutral
model with purifying selection) and a three-rate model (M2a) allowing for some sites to
evolve under positive selection with a pN/pS ratio above one (positive selection model) (Yang
2007). The best fitting model according to a likelihood ratio test (False Discovery Rate < 0.01)
was kept for further analyses. Such model-based inference specifically accounts for
heterogeneous selective pressures within genes, allowing inference of pN/pS ratios for sites under purifying selection and for sites under positive selection, if any.

The average heterozygosity computation assumes an infinite site model and does not allow for recurrent mutations. The model-based approach, while accounting for recurrent mutations and more accurate mutational models, assumes an underlying tree structure and no recombination in the region under study (Anisimova et al. 2001). As Z. tritici exhibits both high mutation rates and intragenic recombination, it is a priori unclear which method provides the best estimates of pN/pS ratios and purifying selection. To assess the sensitivity of model-based estimates in the presence of recombination, we used coalescent simulations (See Methods). We find that estimates of purifying selection based on the pN/pS ratio (denoted “ω0” in the M1a and M2a models) are indeed inflated in the presence of high recombination rates (Figure 3). While the effect is significant, it is of small size. Furthermore, it goes in the opposite direction of what is expected under background selection, as selection should be more efficient in highly recombining regions and lead to the removal of more deleterious mutations reflected by a lower omega ω0 value. We therefore conclude that the model-based inference of purifying selection is conservative and we used these estimates in the further analyses. We estimated the strength of purifying selection for a gene as the ω0 parameter of the best fitting model between M1a and M2a. This procedure ensures that we only measure the strength of purifying selection, even in the presence of positions evolving under positive selection.

To investigate which genome parameters determine the strength of purifying selection in Z. tritici, we compared ω0 to 1) mean gene expression, 2) recombination rate in flanking non-coding regions given by the parameter ρ, 3) gene length, 4) local gene density, 5) local density of transposable elements (TE) and 6) GC content at third codon positions. We further considered whether the gene 7) is predicted to be secreted, 8) is located on an accessory chromosome and 9) is an orphan gene. We fitted a linear model with ω0 as a response variable and the above-mentioned observed statistics as explanatory variables. We identify several variables, which significantly explain the strength of purifying selection (Table 2): the length of the coding protein (number of codons) has a highly significant negative effect, meaning that evolution of larger genes is more constrained. We also observe that orphan genes, which in general are shorter than core genes (mean 325 amino acids vs 516 amino acids on average, Wilcoxon test with p-value < 2.2e-16), are less conserved. The interaction between these two factors, significant at the 5% level, is negative. This implies that the correlation between ω0 and gene length is even more negative in orphan genes than in core
genes. The single effect of orphan genes is highly significant and positive, which suggests that orphan genes are less constrained than core genes of the same length (Table 2). Many orphan genes in the *Z. tritici* genome have likely evolved *de novo* and may not encode proteins with essential functions. Our observation of higher ω₀ in orphan genes is consistent with findings in other organisms, e.g. in primates where younger lineage specific genes evolve under less selective constraints than older “common” primate genes (Cai and Petrov 2010). Other studies that document an impact of gene length on evolutionary rates likewise show that smaller genes evolve faster than long genes (Liao et al. 2006).

We used *Z. tritici* gene expression data from early host colonization (four days after spore inoculation on leaves of seedlings of a susceptible wheat host) and *in vitro* growth to assess the correlation of transcription level and gene evolution (Kellner et al. 2014). Studies in other organisms including yeast and bacteria have documented a strong impact of expression levels on gene evolution and find that highly expressed genes are more conserved (Drummond et al. 2006; Liao et al. 2006). Consistent with these observations we also find a significant negative correlation between ω₀ and gene expression (Kendall’s τ = -0.0759, p-value < 2.2e-16) showing that highly expressed genes are more conserved in *Z. tritici*. However, interestingly, the effect of gene expression is non-significant when accounting for the other explanatory variables in the linear model, which indicates that at least part of this correlation is due to confounding factors correlating with both ω₀ and gene expression.

We furthermore find a highly significant negative effect of recombination rate on ω₀, implying that purifying selection is more efficient in regions with high recombination rate (Figure 4). Also, we show a significant interaction with local gene density (Figure 4). According to the coefficient of our linear model (Table 2), ω₀ can be modelled as:

\[
ω₀ \sim -0.24ρ - 0.94p^*g
\]

\[
ω₀ \sim -(0.24 + 0.94g)ρ
\]

where ρ is the population recombination rate and g is the proportion of coding sites in the region. These results are in agreement with a model of background selection, where purifying selection at linked loci reduces the local effective population size, therefore reducing the efficacy of selection and allowing slightly deleterious mutations to spread more frequently than at unlinked loci (Charlesworth et al. 1993). The effect of background selection is expected to increase with the number of sites at which deleterious mutations can occur (d, assumed to be proportional to the number of coding sites), and to decrease with recombination rate (r), which breaks linkage disequilibrium. The theoretical prediction of the reduction of diversity due to background selection has the form \( \exp(d/r) \), which implies that the effect of recombination on background selection is stronger in regions with higher gene
densities (Nordborg et al. 1996). This finding supports a central role of recombination in the
removal of non-adaptive mutations in the genome of *Z. tritici*. A similar correlation was
shown in *Drosophila melanogaster* where the strength of selection was measured in high and
low recombination genomic regions (Campos et al. 2014).

Lastly, we find that GC content is significantly negatively correlated with \( \omega_0 \), that is, genes
with higher GC tend to be under stronger purifying selection. This may be explained by an
indirect effect of recombination as the GC content at third codon positions (GC3) and
recombination rate are positively correlated (Kendall’s \( \tau = 0.097 \), \( p \)-value < 2.2e16). A
similar correlation of recombination and GC3 is found in other organisms (Duret 2002). In
*Saccharomyces cerevisiae* this correlation has been explained by the impact of biased gene
conversion on sequences evolution (Birdsell 2002). However, a thorough search for signatures
of GC-biased gene conversion did not find any pervasive effect of this phenomenon in *Z. tritici*
(Stukenbrock and Dutheil 2017). An alternative explanation is selection for codon usage. In *Z.
tritici*, as in several other organisms, preferred codons often contain a cytosine or a guanine
at their third position (Figure S2) (Duret and Mouchiroud 1999). As highly expressed genes are
more conserved and enriched in preferred codons, the relationship between \( \omega_0 \) and GC
might be a by-product of the correlation with gene expression.

Detection of on-going diversifying selection in *Z. tritici*

Next we aimed at detecting sites in the *Z. tritici* genome showing signatures of positive
diversifying selection. We fitted models of codon sequence evolution as implemented in the
CODEML program of the PAML package to detect genes with significant signatures of positive
selection using likelihood ratio tests (Yang 2007). Two models are typically compared: the
“null” model M1a, where sites evolve only under neutral or purifying selection (\( \omega \leq 1 \), and
an alternative model M2a where sites in coding sequence also can evolve with a ratio of non-
synonymous over synonymous mutations above one (\( \omega > 1 \)). When applied to population
data, sites with \( \omega > 1 \) can reflect positive diversifying selection (Anisimova et al. 2001).

We fitted codon models for genes present in at least three isolates (83% of genes located on
core chromosomes and 31% of genes on the accessory chromosomes) (Table S3). After
correcting for multiple testing, we identified a final set of 786 genes (including 24 on the
accessory chromosomes) having sites evolving under positive diversifying selection (False
Discovery Rate < 0.01) (Table S3). In total, our analyses recognized 4,686 individual codons
evolving under positive selection in the *Z. tritici* genome (0.001% of 4,640,090 analyzed
codons). In several pathogen genomes, rapidly evolving genes are found clustered in
particular genomic environments often associated with repetitive sequences (e.g. (Raffaele et al. 2010; Dutheil et al. 2016)). To address if the same pattern is found in Z. tritici, we assessed the spatial distribution of genes under positive selection along chromosomes (see Materials and Methods). Of the 786 positively selected genes, 240 are located within a distance of 5 kb from each other and thereby form 108 clusters containing two to four genes with signatures of positive selection. At the genome scale, clusters containing two or three genes do not show a significant pattern as the same pattern can be obtained by randomly distributing the positively selected genes across the genome. However, there are two significant clusters (p-value = 1.8e-3) containing at least four genes. One of the clusters comprising four genes is located in a 24kb region of chromosome 5. The other cluster comprising eight genes is located in a 31 kb region of chromosome 9. For genes in both clusters no functional relevance can be assigned. Aside these two clusters, there is no evidence that positively selected genes are physically associated with each other in Z. tritici, nor that they are significantly enriched or depleted in repeat rich regions.

The molecular battlefield between plants and their pathogens involves proteins secreted from both host and pathogen cells (Lo Presti et al. 2015). Pathogenicity related genes therefore often encode signal peptides targeting the translated protein for secretion. As shown above, adaptive substitutions are fixed at a much higher rate in candidate effectors than in non-effector genes. This observation is in agreement with an antagonistic co-evolution of genes involved in host-pathogen protein-protein interactions. We also asked whether positive diversifying selection, consistent with a trench-warfare scenario of co-evolution, drives effector evolution in Z. tritici. To this end we assessed the proportion of effector candidates among the genes predicted to be under diversifying selection, but found no evidence for a particular enrichment of these genes ($X^2$ test, p-value = 0.171).

We further extended our analyses of the genes predicted to be under positive diversifying selection by characterizing known proteins domains. From the 786 candidate genes under positive selection, 602 of the encoded proteins (76.5%) have an in silico attributed function or harbour known protein domains. We conducted a PFAM domain enrichment analyses and identified 21 significantly enriched domains (FDR <= 0.05) (Table S4). These domains can be grouped into different categories based on their associated molecular function and include carbohydrate-active enzymes (CAZymes), polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), and cellular transporters (Table S4). Several of these categories are relevant for the pathogen to interact with its host. For example, CAZymes are proteins involved in the break down of glycosidic bonds contained in plant cell walls (André et al. 2010; Dutheil et al. 2016).
and PKS and NRPS are multimodular enzymes involved in the biosynthesis of secondary metabolites of which many also have been shown to be involved in virulence of plant pathogenic fungi (reviewed in (Howlett 2006)).

Among the positively selected Z. tritici genes, we also searched for homologs of known virulence factors described in other plant pathogens. The PFAM domain PF14856 corresponds to the mature part of a virulence factor Ecp2 described in the tomato fungal pathogen *Cladosporium fulvum* (Van den Ackerveken et al. 1993). Ecp2 has been described in several other plant pathogens (Stergiopoulos et al. 2010) and has three homologs in Z. tritici. We find that two of these homologs comprise sites under positive selection supporting a virulence related role of this gene also in this wheat pathogen.

**Transposable elements and the origin of new adaptive mutation**

In some filamentous Ascomycetes, the phenomenon Repeat Induced Point (RIP) mutations is a genome defence mechanism that targets repetitive DNA. While experimental evidence for RIP so far only has been documented in a few species (Cambareri et al. 1989; Idnurm and Howlett 2003), sequence signatures of RIP are reported in several other Ascomycete and Basidiomycete species (Hood et al. 2005; Laurie et al. 2012). During meiosis RIP generates single nucleotide changes mutating cytosine bases to thymine with a preferential bias towards CpA dinucleotides (Cambareri et al. 1989). The accelerated mutation rate in repetitive DNA or duplicated sequences mediated by RIP can interfere with the evolution of non-repetitive DNA via “leakage” of RIP mutations into neighboring genes (Daverdin et al. 2012). To infer the potential impact of RIP-like mutations on gene evolution in Z. tritici we assessed the proportion of CpA → TpA (and GpT → ApT for the reverse strand) mutations within all dinucleotide substitutions using a substitution mapping approach for genes located at a maximum distance of 500, 1,000, 1,500 and 2,000 bp from a TE. We find a significant enrichment of RIP-associated dinucleotide substitutions in genes located close to TEs suggesting “leakage” of RIP mutations (Table S5). The proportions of RIP mutations decrease with the distance to TEs supporting that the CpA → TpA mutations indeed originate from RIP. However, we find no enrichment of positively selected sites among the RIP-associated mutations, suggesting that in Z. tritici, RIP is not a major driver of adaptive evolution.

**Signatures of past selection in Z. tritici and related species**

In a previous study the evolutionary history of Z. tritici was inferred using a whole genome coalescence analyses (Stukenbrock et al. 2011). We showed that divergence of Z. tritici and its
sister species *Z. pseudotritici* occurred recently and likely coincides with the onset of wheat domestication and thereby specialization of *Z. tritici* to a new *Triticum* (wheat) host. We hypothesize that genes important for the colonization of distinct hosts have been under selection during the divergence of *Zymoseptoria* species. To infer signatures of past selection we applied the branch model implemented in the program package PAML to estimate the branch-specific dN/dS ratio for core *Zymoseptoria* genes (present in all four species *Z. tritici, Z. pseudotritici, Z. ardabiliae* and *Z. brevis* (Grandaubert et al. 2015)) (Yang 2007). The branch-specific dN/dS ratios reflect the proportion of non-synonymous to synonymous substitutions accumulated in each branch of the *Zymoseptoria* phylogeny. Our analyses identified 47 genes with a dN/dS ratio > 1 on the *Z. tritici* branch indicative of increased non-synonymous divergence, 54 genes in *Z. pseudotritici*, 60 genes in *Z. brevis* and 15 genes in *Z. ardabiliae* (Figure 5). Based on their putative function in host-pathogen interactions, we hypothesized that some positively selected genes encode secreted proteins and putative effectors. Indeed we find a significant enrichment of putative effectors among the positively selected genes on the *Z. tritici* branch (10 genes encoding effector candidates, $\chi^2$ test, $p = 5.272e-05$), and the *Z. pseudotritici* branch (9 genes encoding effector candidates, $\chi^2$ test, $p = 0.00315$). For *Z. ardabiliae* and *Z. brevis* the branch-specific positively selected genes were however not enriched in effector candidates (in *Z. ardabiliae* 1 gene encoding an effector candidate, $\chi^2$ test, $p=0.768$, and in *Z. brevis* 8 genes encoding effector candidates, $\chi^2$ test, $p=0.1165$). We next performed a PFAM domain enrichment analysis for the positively selected genes with a predicted function to address if some functional domains are significantly enriched in this set of genes (FDR < 0.05) (Table S6). The majority of genes encode proteins of unknown function, however among the functionally characterized proteins we find a gene encoding a regulator of chromosome condensation that we previously described to be functionally relevant for virulence in *Z. tritici* (Poppe et al. 2015b). Also our analyses reveal an enrichment of genes encoding proteinases, one gene encoding Lysin motifs (LysM) already described as an effector in *Z. tritici* and other fungal pathogens (de Jonge et al. 2010; Marshall et al. 2011), and one gene encoding a CFEM domain. CFEM domains have been described in G-protein-coupled receptors in the rice blast pathogen *Magnaporthe oryzae* playing an important role in pathogenicity (Kulkarni et al. 2005). These genes provide interesting candidates for further studies of molecular determinants of host specificity in *Z. tritici*. Altogether, our results provide evidence for both arms race and trench warfare evolution acting on the genome of *Z. tritici*, yet involving distinct categories of genes.
Conclusion

By analysing the assembled genomes of thirteen isolates of the fungus *Z. tritici* we provided a detailed analysis of adaptive evolution and the underlying determinants of adaptation in the wheat pathogen. An exceptionally high level of structural variability characterizes the genome of *Z. tritici*, and we here identify more than one thousand previously uncharacterized genes on accessory chromosomes, presenting presence-absence polymorphism between isolates. We hypothesise that this structural variation correlates with phenotypic variation that translates into different virulence phenotypes of the pathogen.

We estimated the genome-wide proportion of non-synonymous substitutions that are due to positive selection using an extension of the McDonald-Kreitmann test. Our analyses revealed a pervasive effect of adaptation on protein evolution, in particular in genes encoding secreted protein and putative virulence determinants. We also find evidence for balancing selection in more than 700 genes showing that both arms race and trench-warfare evolution drive adaptive evolution in *Z. tritici*. Genes evolving under balancing selection are not enriched with effector genes, we find an enrichment of other pathogenicity related genes including enzymes known to be involved in secondary metabolism and pathogenic lifestyles. Maintaining functional variation in these enzymes may be important for the fungus to interact with a diversity of host produced compounds.

While positive selection and rapid adaptation is important to successfully overcome host defences or avoid host recognition, purifying selection is essential to conserve protein function for all basic processes including growth, metabolism, and reproduction. We report a strong impact of recombination on the strength of purifying selection, suggesting a pervasive signature of background selection and highlighting the role of sexual reproduction in the adaptation of fungal pathogens.

Lastly we identify past signatures of selection by comparison of four different *Zymoseptoria* species. These points to a small number of genes containing a significant excess of species-specific adaptive substitutions in each of four closely related species.

In conclusion, detailed analyses of genomic variation in this important wheat pathogen reveal a strong potential for rapid adaptation mediated by a high efficacy of selection in particular affecting virulence-related genes.

Materials and Methods
Re-sequencing, assembly and alignment of *Z. tritici* isolates

In this study we used a geographical collection of thirteen field isolates of *Z. tritici* isolated from infected leaves of bread wheat (*Triticum aestivum*) (Table S1). Genome data of three isolates, including the reference isolate IPO323, were published previously (Goodwin et al. 2011; Stukenbrock et al. 2011). For the remaining ten isolates full genomes were sequenced. DNA extraction was performed as previously described (Stukenbrock et al. 2011). Library preparation and paired end sequencing using an Illumina HiSeq2000 platform were conducted at Aros, Skejby, Denmark. Sequence data of the ten isolates has been deposited under the NCBI BioProject IDs PRJNA312067. We used SOAPdenovo2 [30] to construct de novo genome assemblies for each isolate independently. For each genome, the k-mer value maximizing the weighted median (N50) of contigs and scaffolds was selected.

Prior to generating a multiple genome alignment, we pre-processed the individual genomes of the thirteen *Z. tritici* isolates. First, we masked repetitive sequences using a library of 497 repeat families identified de novo in three *Zymoseptoria* species [19]. Repeats were soft-masked using the program RepeatMasker (option -xsmall) to retained information of repeat sites in the alignment (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at [http://repeatmasker.org](http://repeatmasker.org)). Second, we filtered the genome assemblies to contain only contigs with a length >= 1 kb. Multiple genome alignments were generated by the MULTIZ program using the LASTZ pairwise aligner from the Threaded Blockset Aligner (TBA) package (Blanchette et al. 2004). The alignment was projected on the IPO323 reference genome using the maf_project program from the TBA package.

Inference of population structure

In order to infer population structure, we generated genealogies of the thirteen isolates using the multiple genome alignment. We used the MafFilter program (Dutheil et al. 2014) to compute pairwise distance matrices using maximum likelihood under a Kimura 2 parameter model in 10 kb sliding windows along the chromosomes of the reference genome. For each window, a BioNJ tree was reconstructed from the distance matrices. The resulting 1,850 genealogies were used to build a super tree using SDM for the generation of a distance supermatrix (Criscuolo et al. 2006) and FastME was used to infer a consensus tree (Lefort et al. 2015). We used the program ADMIXTURE (Alexander et al. 2009) to estimate the number of ancestral populations based on single nucleotide polymorphism data. Input files for ADMIXTURE were generated from the filtered alignment using MafFilter (Dutheil et al. 2014).
Presence-absence patterns of chromosomes and genes

Based on the multiple genome alignment, the numbers of aligned bases for each genome were computed in 10 kb non-overlapping sliding windows along the alignment projected on the reference sequence. To assess the presence-absence of protein coding genes we used a recently updated annotation of the reference genome of *Z. tritici* IPO323 (Grandaubert et al. 2015). Nucleotide sequences corresponding to genes were used as queries for BLASTn searches against each of the re-sequenced genomes. The presence of each gene in the individual genomes was assessed if the best BLAST hit showed at least 90% (and up to 110%) of coverage of the total length of the IPO323 gene.

Estimation of rates of adaptation

Based on the coordinates of each predicted gene model in the reference genome IPO323 (Goodwin et al. 2011; Grandaubert et al. 2015), exons were extracted from the multiple genome alignment of *Z. tritici* isolates using MaffFilter (Dutheil et al. 2014). Complete coding sequences (CDS) were concatenated to generate individual alignments of all orthologous CDS. If one or more exons were not extracted from the alignment due to missing information, the gene was discarded from further analyses. Each complete CDS alignment was filtered according to the following criteria: (i) CDS were discarded if they contained more than 5% gaps in one or more individuals, (ii) CDS with premature stop codon were likewise deleted, and (iii) only alignments comprising three or more CDS were kept. In some cases, due to indels in the genome alignment, the codon phasing of some genes was lost. This issue was overcome by refining the CDS alignment using the codon-based multiple alignment program MACSE (Ranwez et al. 2011). The final data set contains 7,040 gene alignments. We further created a data set containing an outgroup sequence, taken from *Z. ardabiliae*, leading to 6,767 alignments. The data set with outgroup was used to infer the synonymous and non-synonymous divergence based on the rate of synonymous and non-synonymous substitutions. The synonymous and non-synonymous unfolded site frequency spectra (SFS) were also computed, using the outgroup sequence to orient mutations. All calculations were performed using the BppPopStats program from the Bio++ Program Suite (Guéguen et al. 2013). We used the dfem program in order to estimate the distribution of fitness effects from the SFS and compute a genome wide estimate of \( \lambda \) and \( \omega_a \), the proportion of mutations fixed by selection and the rate of adaptive substitutions respectively (Galtier 2016). The following models were fitted and compared using Akaike’s information criterion: Neutral, Gamma Gamma-Exponential, Displaced Gamma, Scaled Beta and Bessel K. Analyses were conducted.
on the complete set of gene alignments, as well as on a subset only on genes encoding effector candidates only. We further stratified our data set according to the flanking recombination rate of each gene in 500 bp 5′ and 3′ flanking regions (Stukenbrock and Dutheil 2017). We discretized the observed distribution of the population recombination rate ρ in 15 categories with similar number of genes, each category comprising between 676 and 749 genes. For each of the 15 gene sets, 10 bootstrap replicates were generated by sampling genes randomly in each category. Genes in each replicate were concatenated and the Dfem program run with the GammaExpo distribution of fitness effect (Galtier 2016). For each recombination category, the mean estimates of α and ω, as well as the standard error over the 10 replicates, were computed and plotted (Figure 2). Kendall correlation tests were performed on the 15 times 10 estimates in order to correlate both α and ω, with ρ.

Model of codon sequence evolution and candidate genes under positive diversifying selection

We used all filtered CDS alignments to reconstruct genealogies for the individual genes using PhyML (model HKY85) (Guindon and Gascuel 2003). To investigate patterns of selection and infer the role of positive selection on adaptive gene evolution, the program CODEML from the PAML package was used (Yang 2007) with the filtered multiple CDS alignments and the corresponding phylogenetic trees as inputs. CODEML allows inference of selection and evolutionary rates by calculating the parameter ω, the ratio of non-synonymous to synonymous rates (dN/dS) for each gene. More specifically, we compared site models that allow ω to vary among codons in the protein (Nielsen and Yang 1998). The models used in this study include the nearly neutral (M1a), positive selection (M2a), beta (M7), and beta&ω (M8) models. A likelihood ratio test (LRT) was used to compare the fit of null models and alternative models, and the significance of the LRT statistic was determined using a χ² distribution. The first LRT tests for the occurrence of sites under positive selection by comparing the M1a and M2a models. In the model M1a sites can be under purifying selection (0 < ω < 1) and evolve by neutral evolution (ω = 0) while the M2a model allows for some sites to be under positive selection (ω > 1). The second LRT compares the M7 and M8 models, where in M7 a discretized beta distribution for ω (limited to the interval [0,1]) over all sites is assumed, while the model M8 allows sites with ω > 1. By allowing for a wider range of strength of purifying selection, the M7 and M8 models are more biologically realistic. The M7-M8 LRT, however, was shown to display an increased false discovery rate compared to the M1a-M2a comparison (Anisimova et al. 2001). For these two tests, p-values were obtained
with a χ² test (df = 2). We corrected for multiple testing and a false discovery rate of 1% was used for the detection of genes under positive selection (Benjamini and Hochberg 1995). Only genes significant for both tests were considered as genes evolving under positive selection.

To address divergent adaptation, we compared gene evolution among four closely related Zymoseptoria species. In a previous study we defined the core proteome of Z. tritici, Z. ardabiliae, Z. brevis and Z. pseudotritici comprising 7,786 orthologous genes (Grandaubert et al. 2015). We generated alignments of the corresponding coding sequences using the MACSE sequence aligner (Ranwez et al. 2011) and used CODEML with a branch model that allows ω to vary among branches of the phylogeny (Nielsen and Yang 1998). As input we applied a non-rooted tree of the four Zymoseptoria species as published in (Stukenbrock et al. 2012). Branch lengths were re-estimated for each gene by CODEML (Yang 2007).

Functional enrichment analysis and gene cluster analysis
PFAM domains were extracted from Interproscan results from (Grandaubert et al. 2015). Only domain hits with e-values lower than 1e-5 were considered resulting in 10,026 domains present in 7,343 genes. Enrichment tests were performed based on contingency tables, counting the number of genes containing the domain and the number of genes which do not contain it, for both the complete proteome and a given set of candidates to test. A χ² test was performed to assess significance.

To analyse the distribution of genes under positive selection, we considered two genes separated by less than 5,000 bp to be clustered and assessed the probability of such clusters under a random distribution of genes along the chromosomes. To do so on a genome-wide scale, we calculated the probability to obtain clusters encompassing from two to ten genes under positive selection when these genes are randomly distributed across all gene coordinates. Based on 10,000 random permutations, it appeared that only clusters containing more than three genes were significant at the 5% level.

Genome-wide analysis of selection patterns
We compared the selection regime of each gene to several variables. Local recombination rate was computed using the fine-scale recombination map of Stukenbrock and Dutheil (Stukenbrock and Dutheil 2017) as the average in the 500 bp 5’ and 3’ flanking regions. Local gene density was computed as the proportion of coding sites in 10 kb upstream and downstream regions. Similarly, the density of TEs was computed as the proportion of sites annotated as TE in a windows starting 10 kb upstream the gene and ending 10 kb
downstream the gene, allowing for TE located within the gene itself. GC content at third
codon position (GC3) was also computed. Expression levels were calculated as in (Kellner et
al. 2014). Mean expression level was computed as the maximum value observed for the gene
in axenic culture or plant infection, each averaged over three biological replicates. Genes
encoding secreted proteins were predicted by the program SignalP (Petersen et al. 2011).
Genes with no homolog in any other species were considered as “orphan” genes and genes
located on accessory chromosomes were considered as “dispensable”.
A linear model was defined with the strength of purifying selection (\( \omega_0 \)) as a response
variable, together with all individual explanatory variables. We further included an interaction
of gene density and recombination rate, as expected under a background selection model and
an interaction between protein length given that orphan genes are reported to be
significantly shorter than other genes (Liao et al. 2006; Cai and Petrov 2010). Genes with \( \omega_0 < 
0.001 \) or \( \omega_0 > 999.0 \) were discarded for further analyses as these estimates most likely
indicate model fitting issues, possibly due to lack of data. A BoxCox transform was used to
normalize the model residues. After normalization, a significant departure to normality was
still observed (Shapiro-Wilk normality test, \( p \)-value = 2.315), but no significant
heteroscedasticity (Harrison-McCabe test, \( p \)-value = 0.102), and model residues were found
to be independent (Box-Ljung test, \( p \)-value = 0.9023). We used the rms package in R to
compute robust regression estimates (Harrell 2015).
We selected the 10% \( Z. \ tritici \) most expressed genes and computed the relative synonymous
codon usage of every codons (Sharp et al. 1986). Analyses were conducted using the 'uco'
function of the seqinr package for R (R Core Team 2013).

**Simulation with recombination**

We used the MS program in order to simulated ancestral recombination graphs (ARGs) for
thirteen individuals and a given recombination rate (Hudson 2002). The resulting ARGs were
used to simulate codon alignments using the bppSeqGen program (Dutheil and Boussau
2008). Codon sites were simulated with dN/dS values of 0.1, 0.3, 0.6, 1.0 or 3.0, with equal
probability. In this set-up, recombination was effectively allowed only between codons and
not within codons. The resulting alignments were given as input to the PhyML program
(Guindon et al. 2005) with a HKY85 substitution model and a discretized gamma distribution
of site-specific rates with 4 classes, as used for inference of gene data sets in this study. M1a
and M2a models were then fitted using the estimated tree as input with CODEML.
Recombination rate of \( p = 0.0, 0.001, 0.01, 0.05 \) and 0.1, for gene lengths of 100, 500, 1,000
and 2,000 codons with 30 independent replicates for each combination of parameters. CODEML output was parsed using BioPython (Cock et al. 2009).

Substitution mapping

In order to assess the putative occurrence of RIP mutations among mutations observed in a given gene, we inferred the amount of CpA → TpA and GpT → ApT dinucleotide substitutions using probabilistic substitution mapping with a dinucleotide substitution model previously fitted to the data using maximum likelihood (Dutheil and Boussau 2008). Given a substitution model and a phylogenetic tree, the probabilistic substitution mapping procedure enabled us to count each type of substitution for a site of an alignment and each branch of a phylogeny, while accounting for the uncertainty of the ancestral sequences at the inner nodes of the phylogeny.

Acknowledgements

The authors thank Nicolas Galtier and Thomas Bataillon for helpful discussion. The study was funded by a Max Planck fellowship and a personal grant from the State of Schleswig-Holstein, Germany both to EHS.

References


BANKE S, MCDONALD BA. 2005. Migration patterns among global populations of the


Daverdin G, Rouxel T, Gout L, Aubertot J-N, Fudal I, Meyer M, Parlarge F, Carpezat J,
Balesdent M-H. 2012. Genome Structure and Reproductive Behaviour Influence the
Evolutionary Potential of a Fungal Phytopathogen. PLOS Pathog. 8:e1003020.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast


Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon
4487.


multiple genome alignment files processor. BMC Genomics 15:53.

Dutheil JY, Mannhaupt G, Schweizer G, Sieber CMK, Münsterkötter M, Güldener U, Schirawski


Rev Genet 8:610–618.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size

Goodwin SB, M’barek S Ben, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, Foster AJ, Van der
pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome

Gossmann Tl, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov M V, Filatov DA, Eyre-
Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in

Grandaubert J, Bhattacharyya A, Stukenbrock EH. 2015. RNA-seq Based Gene Annotation and
Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria
Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements.

G3 Genes | Genomes | Genetics 5(7):1323-33


Harrell F. 2015. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer


Idnurm A, Howlett BJ. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete Leptosphaeria maculans.


Klosterman SJ, Subbarao K V, Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B,


Möller M, Stukenbrock EH. 2017. Evolution and genome architecture in fungal plant


R Core Team. 2013. R: A language and environment for statistical computing.


Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, Freitag M,


Tables

Table 1: Estimates of the proportion of adaptive mutation (α) under various models of distribution of fitness effects with the DFEM method. Each line corresponds to one model, with its corresponding number of parameters, log likelihood and Akaike’s information criterion (AIC), as well as the inferred proportion, α, and the rate, ω_a, of adaptive mutations. The upper part of the table corresponds to models fitted on all genes, while the lower part corresponds to model fitted on predicted effector genes only.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Model</th>
<th>Nb. parameters</th>
<th>Log likelihood</th>
<th>AIC</th>
<th>alpha</th>
<th>omega_a</th>
</tr>
</thead>
<tbody>
<tr>
<td>All genes</td>
<td>Neutral</td>
<td>16</td>
<td>-1673,86</td>
<td>3379,72</td>
<td>0,26</td>
<td>0,03</td>
</tr>
<tr>
<td>All genes</td>
<td>Gamma</td>
<td>17</td>
<td>-256,14</td>
<td>546,28</td>
<td>0,5</td>
<td>0,05</td>
</tr>
<tr>
<td>All genes</td>
<td>GammaExpo</td>
<td>19</td>
<td>-237,73</td>
<td>513,46</td>
<td>0,44</td>
<td>0,05</td>
</tr>
<tr>
<td>All genes</td>
<td>DisplacedGamma</td>
<td>18</td>
<td>-256,14</td>
<td>548,28</td>
<td>0,5</td>
<td>0,05</td>
</tr>
<tr>
<td>All genes</td>
<td>ScaledBeta</td>
<td>18</td>
<td>-303,8</td>
<td>643,6</td>
<td>0,52</td>
<td>0,05</td>
</tr>
<tr>
<td>All genes</td>
<td>BesselK</td>
<td>19</td>
<td>-294,94</td>
<td>627,88</td>
<td>0,42</td>
<td>0,04</td>
</tr>
<tr>
<td>SSP</td>
<td>Neutral</td>
<td>16</td>
<td>-102,64</td>
<td>237,28</td>
<td>0,32</td>
<td>0,07</td>
</tr>
<tr>
<td>SSP</td>
<td>Gamma</td>
<td>17</td>
<td>-90,05</td>
<td>214,1</td>
<td>0,52</td>
<td>0,11</td>
</tr>
<tr>
<td>SSP</td>
<td>GammaExpo</td>
<td>19</td>
<td>-84,6</td>
<td>207,2</td>
<td>0,68</td>
<td>0,14</td>
</tr>
<tr>
<td>SSP</td>
<td>DisplacedGamma</td>
<td>18</td>
<td>-90,41</td>
<td>216,82</td>
<td>0,53</td>
<td>0,11</td>
</tr>
<tr>
<td>SSP</td>
<td>ScaledBeta</td>
<td>18</td>
<td>-86,69</td>
<td>209,38</td>
<td>0,6</td>
<td>0,12</td>
</tr>
<tr>
<td>SSP</td>
<td>BesselK</td>
<td>19</td>
<td>-86,41</td>
<td>210,82</td>
<td>0,5</td>
<td>0,1</td>
</tr>
</tbody>
</table>

Table 2: Coefficient and associated p-values of each variable explaining the distribution of the intensity of purifying selection, using linear modeling.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Coefficient</th>
<th>S.E.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1,5992</td>
<td>0,2323</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Expression</td>
<td>0,0001</td>
<td>0,0001</td>
<td>0,109</td>
</tr>
<tr>
<td>Recombination rate</td>
<td>-0,2419</td>
<td>0,0604</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Protein size</td>
<td>-0,0004</td>
<td>0,0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Gene density</td>
<td>-0,1098</td>
<td>0,091</td>
<td>0,228</td>
</tr>
<tr>
<td>TE density</td>
<td>0,033</td>
<td>0,432</td>
<td>0,939</td>
</tr>
<tr>
<td>GC 3rd codon position</td>
<td>-5,6186</td>
<td>0,3441</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Secreted</td>
<td>0,0796</td>
<td>0,0766</td>
<td>0,2985</td>
</tr>
<tr>
<td>Dispensable</td>
<td>0,4291</td>
<td>0,2514</td>
<td>0,088</td>
</tr>
<tr>
<td>Orphan</td>
<td>0,9212</td>
<td>0,1834</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Recombination rate * Gene density</td>
<td>-0,9378</td>
<td>0,3509</td>
<td>0,0076</td>
</tr>
<tr>
<td>Protein size * Orphan</td>
<td>-0,0012</td>
<td>0,0006</td>
<td>0,0444</td>
</tr>
</tbody>
</table>
Figures

**Figure 1:** Population structure of the thirteen *Z. tritici* isolates. A) Consensus super tree of the thirteen isolates based on 1,850 genealogies estimated in 10 kb sliding windows along the multiple genome alignment. This tree suggests the grouping of the isolates into two populations originating from Europe and Iran. B) Based on SNP data, the program ADMIXTURE estimated that the best separation of the isolates is also in two populations (k=2). However, the use of k=3 highlighted a German sub-population within the European isolates.

**Figure 2:** Estimates of A) the proportion of adaptive substitution $\alpha$, and B) the rate of adaptive substitution, $\omega_a$, as a function of the population recombination rate ($\rho$). Each point and bars
replicates. The blue lines show the fitted linear models.

Figure 3: Impact of recombination on the estimation of the strength of purifying selection using phylogenetic models of codon sequence evolution. Parameters are the ratio of transitions over transversions (kappa), the ratio of the rates of synonymous over non-synonymous mutations for sites evolving under purifying selection (omega0) in the two models M1 and M2, and the ratio of the rates of synonymous over non-synonymous mutations for sites evolving under positive selection (omega1) for model M2 (see methods). Rho: scaled recombination rate used in the simulations. Length: length of the gene used in the simulation. X-axis: average number of distinct topologies per simulated site, increasing with the recombination rate. Y-axis: parameter estimate.
Figure 4: Patterns of selection along the genome of *Z. tritici*. Recombination, intensity of purifying selection (as estimated from the Pn/Ps ratio), density of coding sites (CDS) and density of transposable elements (TE) are plotted in windows of 150 kb along the thirteen essential chromosomes.
Figure 5: Branch-specific signatures of divergent adaptation in *Zymoseptoria* species.

Phylogeny of four *Zymoseptoria* species based on integrating the mean dN / dS ratio (in red) and the number of positively selected genes (in green) estimated from the 7,786 genes shared by the species using the branch model implemented in PAML (Yang 2007).

Supplementary materials

Table S1: Summary table of isolates used in this study and genome assembly statistics.

Table S2: Summary statistics of the multiple genome alignment of thirteen *Z. tritici* genomes.

Table S3: Output of the PAML analysis using codon site models for the 9,412 filtered CDS of *Z. tritici*.

Table S4: Functional enrichment analysis using PFAM domains for the 786 genes with sites under positive selection in *Z. tritici*.

Table S5: Representation of RIP mutations in genes according to their distance to transposable elements.

Table S6: Functional enrichment analysis using PFAM domains for the genes under positive selection in four *Zymoseptoria* species.

Figure S1: Chromosome polymorphism in the twelve re-sequenced genomes compared to the IPO323 reference genome. Based on the multiple genome alignment, for each re-sequenced genomes the alignment coverage was calculated in 10 kb sliding windows along the total length of each chromosome using the genome of IPO323 as a reference. Chromosomes 1 to 13 (in blue) are well conserved among the different isolates and constitute the “core” chromosomes while the chromosomes 14 to 21 (in red) appear to be absent or incomplete at different frequencies.

Figure S2: Codon usage in *Z. tritici*. Relative synonymous codon usage (RSCU) in the 10% most expressed genes of *Z. tritici*. Codon usage, according to the base type at the third position.