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

93 mutations and will therefore accumulate over time as species diverge, resulting in a higher
94 dN/dS ratio (Eyre-Walker 2006; Galtier 2016).
95 The McDonald-Kreitman table forms the basis for the quantification of the proportion of
96 adaptive substitutions in a gene fixed by positive selection defined as $\alpha = 1 - (pN / pS) / (dN /$
97 $dS) = 1 - (dS * pN) / (dN * pS)$ (Smith and Eyre-Walker 2002). Additional modeling of the
98 fitness effect of substitutions allows α to account for the presence of slightly deleterious or
99 slightly advantageous mutations that otherwise can bias the inference of positively selected
100 substitutions (Eyre-Walker and Keightley 2007). These methods have been applied to a range
101 of different species and have been used to show that the rate of adaptive evolution varies
102 largely among different taxa (Eyre-Walker 2006; Galtier 2016). Overall, the variation in α
103 correlates positively with the average effective population size (Ne), where species with large
104 Ne more efficiently remove slightly deleterious mutations resulting in a reduced pN value and
105 lower pN/pS ratios compared to species with small Ne (Galtier 2016).
106 In pathogen genomes, increased rates of adaptation, as occurring by arms-race evolution
107 where new non-synonymous mutations repeatedly are selected for, will result in higher
108 dN/dS ratios. Indeed, evolutionary predictions based on dN/dS analyses and McDonald-
109 Kreitman tests have identified effectors and other pathogenicity-related genes in pathogen
110 genomes (Raffaele et al. 2010; Stukenbrock et al. 2011). However, genes that evolve by
111 trench warfare evolution leave different signatures in the genome sequence. This type of
112 selection can also be referred to as positive diversifying selection where a diversity of alleles
113 at the selected locus is maintained by natural selection. The genic signature of positive
114 diversifying selection is an increased proportion of non-synonymous polymorphisms
115 (Charlesworth 2006). Yet, in a functionally important gene, the majority of non-synonymous
116 sites will be evolving under selective constraints, and only a small number will be allowed to
117 change without negatively affecting protein function. Therefore, in most cases, searching for
118 signatures of positive diversifying selection simply by comparing the average proportion of
119 non-synonymous to synonymous polymorphisms in a gene will show little statistical power
120 (Nielsen and Yang 1998). To circumvent this, evolutionary codon models that allow for
121 varying selection intensity across a gene sequences can be applied to identify specific codon
122 positions that have an excess of non-synonymous to synonymous mutations (a higher pN/pS
123 ratio) (Nielsen and Yang 1998).
124 In this study we assess the different types of selection affecting the genome of a fungal wheat
125 pathogen *Zymoseptoria tritici* (synonym *Mycosphaerella graminicola*). The genome of this
126 pathogen consists of thirteen core chromosomes and several accessory chromosomes that

127 are present at different frequencies in different individuals, constituting a particular case of
128 karyotypic polymorphism (Goodwin et al. 2011). The accessory chromosomes comprise
129 repeat rich, heterochromatic DNA with a low gene content and a so far unknown functional
130 relevance (Grandaubert et al. 2015; Schotanus et al. 2015). We have previously shown that
131 evolutionary rates on these chromosomes are particularly high suggesting that genes on the
132 accessory chromosomes in general evolve under less selective constraints (Stukenbrock et al.
133 2010).
134 Previous studies based on comparative population genomic analyses of *Z. tritici* and two
135 closely related species, *Z. pseudotritici* and *Z. ardabiliae* used genome-wide estimates of non-
136 synonymous and synonymous divergence to identify past species-specific signatures of
137 selection in the wheat pathogen (Stukenbrock et al. 2010; Stukenbrock et al. 2011).
138 Functional characterization of some of these genes revealed amino acid substitutions
139 important for *in planta* development and asexual spore formation in the wheat-adapted
140 pathogen, and thereby confirmed the use of evolutionary predictions to identify functionally
141 relevant traits (Poppe et al. 2015a).
142 While past signatures of selection can be related to ecological specialization of *Zymoseptoria*
143 species, ongoing positive selection likely reflects host-pathogen arms race or trench warfare
144 evolution and recurrent adaptations to changing environmental conditions including fungicide
145 treatments and wheat cultivars. Population genomic studies of *Z. tritici* have revealed that the
146 pathogen has a high effective population size, which correlates positively with a strong
147 efficacy of natural selection, i.e. the ability to rapidly fix advantageous mutations and
148 eliminate non-advantageous (Stukenbrock et al. 2011). Additional parameters that can
149 influence rapid adaptation are recombination and mutation rates, as well as genome
150 architecture. Recombination rate variation has been studied in *Z. tritici* by experimental
151 mating and population genomic sequencing. Independently, these approaches have
152 confirmed exceptionally high recombination rates (~60cM/Mb) ((Croll et al. 2015;
153 Stukenbrock and Dutheil 2017)). So far the implications of such high recombination rates on
154 gene evolution are not known.
155 We apply a population genomics approach to infer genome-wide positive and purifying
156 selection among thirteen isolates of *Z. tritici* collected from bread wheat in Europe and the
157 Middle East. Our analyses based on more than 1.4 million single nucleotide polymorphisms
158 (SNPs) and including 700,000 coding sites allow us to identify past and on-going signatures of
159 selection in the genome of *Z. tritici*. Notably, our analyses reveal a strong importance of
160 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 π = 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

195 pathogens such as the wheat yellow rust fungus *Puccinia striiformis f.sp. tritici* and the wheat
196 scab pathogen *Fusarium graminearum* (O'Donnell et al. 2000; Ali et al. 2014), and shows that
197 local adaptation and population divergence can shape the genetic structure of cereal
198 pathogens even in the presence of global agricultural trade and long distance pathogen
199 dispersal (Brown and Hovmøller 2002).
200
201 ***Z. tritici* shows presence-absence polymorphism of genes and whole chromosomes**
202 *Z. tritici* exhibits dramatic patterns of chromosome length polymorphisms and presence-
203 absence variation of accessory chromosomes (Mehrabi et al. 2007; Brunner et al. 2013;
204 McDonald et al. 2016). Based on the multiple genome alignment, we here investigated
205 presence-absence polymorphisms of chromosomes, chromosome fragments and genes
206 among the thirteen isolates. We focused on non-repetitive DNA and show, in agreement with
207 a previous study (Croll et al. 2013), that 96.4% of the core chromosomes (chromosome 1 to
208 13) are shared among all isolates (Figure S1). However, as expected, the mapping of
209 assembled contigs aligning to the accessory chromosomes varied considerably between
210 isolates (Figure S1).
211 Overall, 9,499 of the 11,839 genes predicted in the reference genome of IPO323 are present
212 in all of the thirteen isolates. Of the remaining, 1,584 genes are located on core chromosomes
213 and 667 on accessory chromosomes. This set of genes comprises genes present at different
214 frequencies among the thirteen *Z. tritici* isolates including 89 genes present only in IPO323
215 and of which 77 locate on core chromosomes.
216 A previous genome comparison of two Swiss *Z. tritici* isolates revealed nearly 300 novel genes
217 not characterized in the reference genome (Plissonneau et al. 2016). We speculated that the
218 different isolates of *Z. tritici* also could contain accessory genes and chromosomes not
219 present in the reference IPO323. To test this we investigated the subset of contigs that were
220 not aligned to the reference genome. We found a total of 5.2 Mb in the twelve re-sequenced
221 *Z. tritici* isolates with no homology to sequences in the reference genome. These new *Z. tritici*
222 sequences, corresponding to accessory fragments, encompass between 602 kb and 995 kb of
223 non-repetitive DNA in each genomes. We applied an *ab initio* gene predictor to assess
224 whether these accessory sequences encode proteins and found evidence for 1,041 novel full-
225 length gene models encoding proteins of at least 30 amino acids. Of these 75% encode
226 proteins without known function consistent with a general pattern observed on the accessory
227 chromosomes with a high proportion of gene models without a predicted function [19].
228 Isolate-specific fragments were previously shown to encode effector candidate genes strongly

229 up-regulated in *Z. tritici* during host infection (Plissonneau et al. 2016). In other fungal
230 pathogens such as *Verticilium dahliae* and *Magnaporthe oryzae* isolate-specific genes were
231 experimentally shown to encode virulence determinants (de Jonge et al. 2012; Xue et al.
232 2012). Based on our findings, we hypothesize that the large complement of isolate-specific
233 genes in thirteen *Z. tritici* genomes could play a role in host-pathogen interactions. Under this
234 scenario, the observed variation in gene content could reflect diversifying selection imposed
235 by genetic traits in a variety of host genotypes.
236
237 **Measures of adaptive evolution in *Z. tritici***
238 We next aimed to obtain a quantitative assessment of adaptive mutations in the genome of *Z.*
239 *tritici*. To this end, we first estimated the non-synonymous and synonymous divergence dN
240 and dS using a genome alignment of *Z. tritici* and its sister species *Z. ardabiliae*. Furthermore,
241 we used the *Z. tritici* SNP data to compute the unfolded site frequency spectra (SFS) of
242 synonymous and non-synonymous sites using *Z. ardabiliae* as an outgroup. By contrasting
243 divergence and polymorphism data, we estimated the parameter α as a measure of the
244 proportion of adaptive non-synonymous substitutions. The SFS is however strongly affected
245 by demography and the presence of slightly deleterious mutations segregating at low
246 frequencies (see review by (Eyre-Walker and Keightley 2007)). State-of-the-art statistical
247 methods account for these effects by modeling the distribution of fitness effects (DFE) of
248 mutations (Gossmann et al. 2010; Galtier 2016). We used the program Dfem to compute α as
249 well as ω_a, the rate of adaptive substitutions, using four distinct DFE models accounting for
250 mutations with both slightly deleterious and beneficial effects (Galtier 2016). We fitted and
251 compared the DFE models (see Materials and Methods) and found that the Gamma-
252 Exponential model fitted our data best (Table 1) in agreement with studies of genetic data
253 from animals (Galtier 2016). This suggests the existence of slightly deleterious, as well as
254 slightly beneficial segregating mutations in the genome of *Z. tritici* (Table 1). These estimates
255 based on a Gamma-Exponential model suggest an α value of 44% as a genome average, and
256 an ω_a value of 0.05. Both values are in the range of what is observed for Mammals (with the
257 exception of Primates) but considerably higher than estimates from plants (Gossmann et al.
258 2010; Galtier 2016). In candidate effector genes, however, the rate of mutations fixed by
259 selection is more than twice as high as the average of all genes (Table 1), with 68% of
260 substitutions being estimated as adaptive. This estimate is close to the highest values
261 reported in animals (Galtier 2016). We further aimed at assessing the impact of
262 recombination on the fixation of adaptive mutations, by considering genes in distinct

263 recombination environments. We report a significant positive correlation between α and ρ
264 (the population scaled recombination rate inferred as ρ = 4Ne x r) (Kendall's tau = 0.25, p-
265 value = 6.816e-6) and ω_a and ρ (Kendall's tau = 0.21, p-value = 1.95e-4, Figure 2), suggesting
266 that a higher recombination rate favors the fixation of adaptive mutation, as expected by a
267 Hill-Robertson effect (Hill and Robertson 1966; Marais and Charlesworth 2003). These results,
268 based on population genomic data of *Z. tritici*, represent the first quantitative assessment of
269 adaptive evolution in the genome of a fungal pathogen. The exceptionally high rate of
270 adaptation in effector genes likely reflects arms race evolution driven by the antagonistic
271 interaction of *Z. tritici* and its host.
272
273 **Strength of purifying selection along the genome of *Z. tritici***
274 Loci at different positions of the genome evolve at different speed. This is due to variation in
275 the effect of different evolutionary processes across the genome as well as variation in the
276 composition of the DNA sequence itself (Ellegren 2014). In sexual organisms, recombination is
277 a strong determinant of evolutionary rates, and differences in the frequency of recombination
278 can strongly impact the rate whereby adaptive mutations can be fixed and whereby (slightly)
279 deleterious mutations are removed (Marais and Charlesworth 2003). Evolutionary processes
280 at any given site in the genome are strongly influenced by selection acting on linked sites, a
281 process referred to as linked selection, including Hill-Robertson interference in the case of
282 linkage with positively selected mutations (Hill and Robertson 1966) and background
283 selection in the case of linkage with mutation under purifying selection (Hill and Robertson
284 1966; Nordborg et al. 1996). Here, we aimed to address the genome-wide strength of
285 purifying selection in *Z. tritici* in protein coding genes, using the ratio of non-synonymous to
286 synonymous polymorphisms (pN/pS ratio).
287 We computed pN and pS for each gene with two methods: (1) by inference of average
288 pairwise heterozigosity (Romiguier et al. 2014; Ellegren and Galtier 2016) and (2) using a
289 maximum likelihood framework and models of sequence evolution (Nielsen and Yang 1998).
290 In this framework, two types of models were compared: a two-rate model (M1a), with sites
291 evolving under neutrality or purifying selection with a pN/pS ratio equal or below one (neutral
292 model with purifying selection) and a three-rate model (M2a) allowing for some sites to
293 evolve under positive selection with a pN/pS ratio above one (positive selection model) (Yang
294 2007). The best fitting model according to a likelihood ratio test (False Discovery Rate < 0.01)
295 was kept for further analyses. Such model-based inference specifically accounts for

296 heterogeneous selective pressures within genes, allowing inference of pN/pS ratios for sites
297 under purifying selection and for sites under positive selection, if any.
298 The average heterozygosity computation assumes an infinite site model and does not allow
299 for recurrent mutations. The model-based approach, while accounting for recurrent
300 mutations and more accurate mutational models, assumes an underlying tree structure and
301 no recombination in the region under study (Anisimova et al. 2001). As *Z. tritici* exhibits both
302 high mutation rates and intragenic recombination, it is a priori unclear which method
303 provides the best estimates of pN/pS ratios and purifying selection. To assess the sensitivity
304 of model-based estimates in the presence of recombination, we used coalescent simulations
305 (See Methods). We find that estimates of purifying selection based on the pN/pS ratio
306 (denoted "ω0" in the M1a and M2a models) are indeed inflated in the presence of high
307 recombination rates (Figure 3). While the effect is significant, it is of small size. Furthermore,
308 it goes in the opposite direction of what is expected under background selection, as selection
309 should be more efficient in highly recombining regions and lead to the removal of more
310 deleterious mutations reflected by a lower omega ω0 value. We therefore conclude that the
311 model-based inference of purifying selection is conservative and we used these estimates in
312 the further analyses. We estimated the strength of purifying selection for a gene as the ω0
313 parameter of the best fitting model between M1a and M2a. This procedure ensures that we
314 only measure the strength of purifying selection, even in the presence of positions evolving
315 under positive selection.
316 To investigate which genome parameters determine the strength of purifying selection in *Z.*
317 *tritici*, we compared ω0 to 1) mean gene expression, 2) recombination rate in flanking non-
318 coding regions given by the parameter ρ, 3) gene length, 4) local gene density, 5) local density
319 of transposable elements (TE) and 6) GC content at third codon positions. We further
320 considered whether the gene 7) is predicted to be secreted, 8) is located on an accessory
321 chromosome and 9) is an orphan gene. We fitted a linear model with ω0 as a response
322 variable and the above-mentioned observed statistics as explanatory variables. We identify
323 several variables, which significantly explain the strength of purifying selection (Table 2): the
324 length of the coding protein (number of codons) has a highly significant negative effect,
325 meaning that evolution of larger genes is more constrained. We also observe that orphan
326 genes, which in general are shorter than core genes (mean 325 amino acids vs 516 amino
327 acids on average, Wilcoxon test with p-value < 2.2e-16), are less conserved. The interaction
328 between these two factors, significant at the 5% level, is negative. This implies that the
329 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 tau = -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:

$$\omega 0 \sim -0.24\rho - 0.94\rho * g$$
$$\omega 0 \sim -(0.24 + 0.94g) * \rho$$

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

366 densities (Nordborg et al. 1996). This finding supports a central role of recombination in the
367 removal of non-adaptive mutations in the genome of *Z. tritici*. A similar correlation was
368 shown in *Drosophila melanogaster* where the strength of selection was measured in high and
369 low recombination genomic regions (Campos et al. 2014).
370 Lastly, we find that GC content is significantly negatively correlated with ω0, that is, genes
371 with higher GC tend to be under stronger purifying selection. This may be explained by an
372 indirect effect of recombination as the GC content at third codon positions (GC3) and
373 recombination rate are positively correlated (Kendall's tau = 0.097, p-value < 2.2e16). A
374 similar correlation of recombination and GC3 is found in other organisms (Duret 2002). In
375 *Saccharomyces cerevisiae* this correlation has been explained by the impact of biased gene
376 conversion on sequences evolution (Birdsell 2002). However, a thorough search for signatures
377 of GC-biased gene conversion did not find any pervasive effect of this phenomenon in *Z. tritici*
378 (Stukenbrock and Dutheil 2017). An alternative explanation is selection for codon usage. In *Z.*
379 *tritici,* as in several other organisms, preferred codons often contain a cytosine or a guanine
380 at their third position (Figure S2) (Duret and Mouchiroud 1999). As highly expressed genes are
381 more conserved and enriched in preferred codons, the relationship between ω0 and GC
382 might be a by-product of the correlation with gene expression.
383
384 **Detection of on-going diversifying selection in *Z. tritici***
385 Next we aimed at detecting sites in the *Z. tritici* genome showing signatures of positive
386 diversifying selection. We fitted models of codon sequence evolution as implemented in the
387 CODEML program of the PAML package to detect genes with significant signatures of positive
388 selection using likelihood ratio tests (Yang 2007). Two models are typically compared: the
389 "null" model M1a, where sites evolve only under neutral or purifying selection (ω <= 1), and
390 an alternative model M2a where sites in coding sequence also can evolve with a ratio of non-
391 synonymous over synonymous mutations above one (ω > 1). When applied to population
392 data, sites with ω > 1 can reflect positive diversifying selection (Anisimova et al. 2001).
393 We fitted codon models for genes present in at least three isolates (83% of genes located on
394 core chromosomes and 31% of genes on the accessory chromosomes) (Table S3). After
395 correcting for multiple testing, we identified a final set of 786 genes (including 24 on the
396 accessory chromosomes) having sites evolving under positive diversifying selection (False
397 Discovery Rate < 0.01) (Table S3). In total, our analyses recognized 4,686 individual codons
398 evolving under positive selection in the *Z. tritici* genome (0,001% of 4,640,090 analyzed
399 codons). In several pathogen genomes, rapidly evolving genes are found clustered in

400 particular genomic environments often associated with repetitive sequences (e.g (Raffaele et
401 al. 2010; Dutheil et al. 2016)). To address if the same pattern is found in *Z. tritici*, we assessed
402 the spatial distribution of genes under positive selection along chromosomes (see Materials
403 and Methods). Of the 786 positively selected genes, 240 are located within a distance of 5 kb
404 from each other and thereby form 108 clusters containing two to four genes with signatures
405 of positive selection. At the genome scale, clusters containing two or three genes do not show
406 a significant pattern as the same pattern can be obtained by randomly distributing the
407 positively selected genes across the genome. However, there are two significant clusters (p-
408 value = 1.8e-3) containing at least four genes. One of the clusters comprising four genes is
409 located in a 24kb region of chromosome 5. The other cluster comprising eight genes is located
410 in a 31 kb region of chromosome 9. For genes in both clusters no functional relevance can be
411 assigned. Aside these two clusters, there is no evidence that positively selected genes are
412 physically associated with each other in *Z. tritici*, nor that they are significantly enriched or
413 depleted in repeat rich regions.
414 The molecular battlefield between plants and their pathogens involves proteins secreted from
415 both host and pathogen cells (Lo Presti et al. 2015). Pathogenicity related genes therefore
416 often encode signal peptides targeting the translated protein for secretion. As shown above,
417 adaptive substitutions are fixed at a much higher rate in candidate effectors than in non-
418 effector genes. This observation is in agreement with an antagonistic co-evolution of genes
419 involved in host-pathogen protein-protein interactions. We also asked whether positive
420 diversifying selection, consistent with a trench-warfare scenario of co-evolution, drives
421 effector evolution in *Z. tritici*. To this end we assessed the proportion of effector candidates
422 among the genes predicted to be under diversifying selection, but found no evidence for a
423 particular enrichment of these genes (X^2 test, p-value = 0.171).
424 We further extended our analyses of the genes predicted to be under positive diversifying
425 selection by characterizing known proteins domains. From the 786 candidate genes under
426 positive selection, 602 of the encoded proteins (76.5%) have an *in silico* attributed function or
427 harbour known protein domains. We conducted a PFAM domain enrichment analyses and
428 identified 21 significantly enriched domains (FDR <= 0.05) (Table S4). These domains can be
429 grouped into different categories based on their associated molecular function and include
430 carbohydrate-active enzymes (CAZymes), polyketide synthases (PKS), non-ribosomal peptide
431 synthetases (NRPS), and cellular transporters (Table S4). Several of these categories are
432 relevant for the pathogen to interact with its host. For example, CAZymes are proteins
433 involved in the break down of glycosidic bonds contained in plant cell walls (André et al.

434 2014), and PKS and NRPS are multimodular enzymes involved in the biosynthesis of
435 secondary metabolites of which many also have been shown to be involved in virulence of
436 plant pathogenic fungi (reviewed in (Howlett 2006)).
437 Among the positively selected *Z. tritici* genes, we also searched for homologs of known
438 virulence factors described in other plant pathogens. The PFAM domain PF14856 corresponds
439 to the mature part of a virulence factor Ecp2 described in the tomato fungal pathogen
440 *Cladosporium fulvum* (Van den Ackerveken et al. 1993). Ecp2 has been described in several
441 other plant pathogens (Stergiopoulos et al. 2010) and has three homologs in *Z. tritici*. We find
442 that two of these homologs comprise sites under positive selection supporting a virulence
443 related role of this gene also in this wheat pathogen.
444
445 **Transposable elements and the origin of new adaptive mutation**
446 In some filamentous Ascomycetes, the phenomenon Repeat Induced Point (RIP) mutations is a
447 genome defence mechanism that targets repetitive DNA. While experimental evidence for RIP
448 so far only has been documented in a few species (Cambareri et al. 1989; Idnurm and Howlett
449 2003), sequence signatures of RIP are reported in several other Ascomycete and
450 Basidiomycete species (Hood et al. 2005; Laurie et al. 2012). During meiosis RIP generates
451 single nucleotide changes mutating cytosine bases to thymine with a preferential bias towards
452 CpA dinucleotides (Cambareri et al. 1989). The accelerated mutation rate in repetitive DNA or
453 duplicated sequences mediated by RIP can interfere with the evolution of non-repetitive DNA
454 via "leakage" of RIP mutations into neighboring genes (Daverdin et al. 2012). To infer the
455 potential impact of RIP-like mutations on gene evolution in *Z. tritici* we assessed the proportion
456 of CpA → TpA (and GpT → ApT for the reverse strand) mutations within all dinucleotide
457 substitutions using a substitution mapping approach for genes located at a maximum distance
458 of 500, 1,000, 1,500 and 2,000 bp from a TE. We find a significant enrichment of RIP-associated
459 dinucleotide substitutions in genes located close to TEs suggesting "leakage" of RIP mutations
460 (Table S5). The proportions of RIP mutations decrease with the distance to TEs supporting that
461 the CpA → TpA mutations indeed originate from RIP. However, we find no enrichment of
462 positively selected sites among the RIP-associated mutations, suggesting that in *Z. tritici*, RIP is
463 not a major driver of adaptive evolution.
464
465 **Signatures of past selection in *Z. tritici* and related species**
466 In a previous study the evolutionary history of *Z. tritici* was inferred using a whole genome
467 coalescence analyses (Stukenbrock et al. 2011). We showed that divergence of *Z. tritici* and its

468 sister species *Z. pseudotritici* occurred recently and likely coincides with the onset of wheat
469 domestication and thereby specialization of *Z. tritici* to a new *Triticum* (wheat) host. We
470 hypothesize that genes important for the colonization of distinct hosts have been under
471 selection during the divergence of *Zymoseptoria* species. To infer signatures of past selection
472 we applied the branch model implemented in the program package PAML to estimate the
473 branch-specific dN/dS ratio for core *Zymoseptoria* genes (present in all four species *Z. tritici*,
474 *Z. pseudotritici*, *Z. ardabiliae* and *Z. brevis* (Grandaubert et al. 2015)) (Yang 2007). The branch-
475 specific dN/dS ratios reflect the proportion of non-synonymous to synonymous substitutions
476 accumulated in each branch of the *Zymoseptoria* phylogeny. Our analyses identified 47 genes
477 with a dN/dS ratio > 1 on the *Z. tritici* branch indicative of increased non-synonymous
478 divergence, 54 genes in *Z. pseudotritici*, 60 genes in *Z. brevis* and 15 genes in *Z. ardabiliae*
479 (Figure 5). Based on their putative function in host-pathogen interactions, we hypothesized
480 that some positively selected genes encode secreted proteins and putative effectors. Indeed
481 we find a significant enrichment of putative effectors among the positively selected genes on
482 the *Z. tritici* branch (10 genes encoding effector candidates, χ^2 test, p = 5.272e-05), and the *Z.*
483 *pseudotritici* branch (9 genes encoding effector candidates, χ^2 test, p = 0.00315). For *Z.*
484 *ardabiliae* and *Z. brevis* the branch-specific positively selected genes were however not
485 enriched in effector candidates (in *Z. ardabiliae* 1 gene encoding an effector candidate,
486 χ^2 test, p=0.768, and in *Z. brevis* 8 genes encoding effector candidates, χ^2 test, p=0.1165). We
487 next performed a PFAM domain enrichment analysis for the positively selected genes with a
488 predicted function to address if some functional domains are significantly enriched in this set
489 of genes (FDR < 0.05) (Table S6). The majority of genes encode proteins of unknown function,
490 however among the functionally characterized proteins we find a gene encoding a regulator
491 of chromosome condensation that we previously described to be functionally relevant for
492 virulence in *Z. tritici* (Poppe et al. 2015b). Also our analyses reveal an enrichment of genes
493 encoding proteinases, one gene encoding Lysin motifs (LysM) already described as an effector
494 in *Z. tritici* and other fungal pathogens (de Jonge et al. 2010; Marshall et al. 2011), and one
495 gene encoding a CFEM domain. CFEM domains have been described in G-protein-coupled
496 receptors in the rice blast pathogen *Magnaporthe oryzae* playing an important role in
497 pathogenicity (Kulkarni et al. 2005). These genes provide interesting candidates for further
498 studies of molecular determinants of host specificity in *Z. tritici*. Altogether, our results
499 provide evidence for both arms race and trench warfare evolution acting on the genome of *Z.*
500 *tritici*, yet involving distinct categories of genes.
501

502

503

504 # Conclusion

505 By analysing the assembled genomes of thirteen isolates of the fungus *Z. tritici* we provided a
506 detailed analysis of adaptive evolution and the underlying determinants of adaptation in the
507 wheat pathogen. An exceptionally high level of structural variability characterizes the genome
508 of *Z. tritici*, and we here identify more than one thousand previously uncharacterized genes
509 on accessory chromosomes, presenting presence-absence polymorphism between isolates.
510 We hypothesise that this structural variation correlates with phenotypic variation that
511 translates into different virulence phenotypes of the pathogen.
512 We estimated the genome-wide proportion of non-synonymous substitutions that are due to
513 positive selection using an extension of the McDonald-Kreitmann test. Our analyses revealed
514 a pervasive effect of adapation on protein evolution, in particular in genes encoding secreted
515 protein and putative virulence determinants. We also find evidence for balancing selection in
516 more than 700 genes showing that both arms race and trench-warfare evolution drive
517 adaptive evolution in *Z. tritici*. Genes evolving under balancing selection are not enriched with
518 effector genes, we find an enrichment of other pathogenicity related genes including
519 enzymes known to be involved in secondary metabolism and pathogenic lifestyles.
520 Maintaining functional variation in these enzymes may be important for the fungus to
521 interact with a diversity of host produced compounds.
522 While positive selection and rapid adaptation is important to successfully overcome host
523 defences or avoid host recognition, purifying selection is essential to conserve protein
524 function for all basic processes including growth, metabolism, and reproduction. We report a
525 strong impact of recombination on the strength of purifying selection, suggesting a pervasive
526 signature of background selection and highlighting the role of sexual reproduction in the
527 adaptation of fungal pathogens.
528 Lastly we identify past signatures of selection by comparison of four different *Zymoseptoria*
529 species. These points to a small number of genes containing a significant excess of species-
530 specific adaptive substitutions in each of four closely related species.
531 In conclusion, detailed analyses of genomic variation in this important wheat pathogen reveal
532 a strong potential for rapid adaptation mediated by a high efficacy of selection in particular
533 affecting virulence-related genes.

534

535 # Materials and Methods

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

537 In this study we used a geographical collection of thirteen field isolates of *Z. tritici* isolated
538 from infected leaves of bread wheat (*Triticum aestivum*) (Table S1). Genome data of three
539 isolates, including the reference isolate IPO323, were published previously (Goodwin et al.
540 2011; Stukenbrock et al. 2011). For the remaining ten isolates full genomes were sequenced.
541 DNA extraction was performed as previously described (Stukenbrock et al. 2011). Library
542 preparation and paired end sequencing using an Illumina HiSeq2000 platform were
543 conducted at Aros, Skejby, Denmark. Sequence data of the ten isolates has been deposited
544 under the NCBI BioProject IDs PRJNA312067. We used SOAPdenovo2 [30] to construct *de*
545 *novo* genome assemblies for each isolate independently. For each genome, the k-mer value
546 maximizing the weighted median (N50) of contigs and scaffolds was selected.

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

557

558 **Inference of population structure**

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

569

570 **Presence-absence patterns of chromosomes and genes**

571 Based on the multiple genome alignment, the numbers of aligned bases for each genome
572 were computed in 10 kb non-overlapping sliding windows along the alignment projected on
573 the reference sequence. To assess the presence-absence of protein coding genes we used a
574 recently updated annotation of the reference genome of *Z. tritici* IPO323 (Grandaubert et al.
575 2015). Nucleotide sequences corresponding to genes were used as queries for BLASTn
576 searches against each of the re-sequenced genomes. The presence of each gene in the
577 individual genomes was assessed if the best BLAST hit showed at least 90% (and up to 110%)
578 of coverage of the total length of the IPO323 gene.

579

580 **Estimation of rates of adaptation**

581 Based on the coordinates of each predicted gene model in the reference genome IPO323
582 (Goodwin et al. 2011; Grandaubert et al. 2015), exons were extracted from the multiple
583 genome alignment of *Z. tritici* isolates using MafFilter (Dutheil et al. 2014). Complete coding
584 sequences (CDS) were concatenated to generate individual alignments of all orthologous CDS.
585 If one or more exons were not extracted from the alignment due to missing information, the
586 gene was discarded from further analyses. Each complete CDS alignment was filtered
587 according to the following criteria: (i) CDS were discarded if they contained more than 5%
588 gaps in one or more individuals, (ii) CDS with premature stop codon were likewise deleted,
589 and (iii) only alignments comprising three or more CDS were kept. In some cases, due to
590 indels in the genome alignment, the codon phasing of some genes was lost. This issue was
591 overcome by refining the CDS alignment using the codon-based multiple alignment program
592 MACSE (Ranwez et al. 2011). The final data set contains 7,040 gene alignments. We further
593 created a data set containing an outgroup sequence, taken from *Z. ardabiliae*, leading to
594 6,767 alignments. The data set with outgroup was used to infer the synonymous and non-
595 synonymous divergence based on the rate of synonymous and non-synonymous
596 substitutions. The synonymous and non-synonymous unfolded site frequency spectra (SFS)
597 were also computed, using the outgroup sequence to orient mutations. All calculations were
598 performed using the BppPopStats program from the Bio++ Program Suite (Guéguen et al.
599 2013). We used the dfem program in order to estimate the distribution of fitness effects from
600 the SFS and compute a genome wide estimate of α and ω_a, the proportion of mutations fixed
601 by selection and the rate of adaptive substitutions respectively (Galtier 2016). The following
602 models were fitted and compared using Akaike's information criterion: Neutral, Gamma
603 Gamma-Exponential, Displaced Gamma, Scaled Beta and Bessel K. Analyses were conducted

604 on the complete set of gene alignments, as well as on a subset only on genes encoding
605 effector canidates only. We further stratified our data set according to the flanking
606 recombination rate of each gene in 500 bp 5' and 3' flanking regions (Stukenbrock and Dutheil
607 2017). We discretized the observed distribution of the population recombination rate ρ in 15
608 categories with similar number of genes, each category comprising between 676 and 749
609 genes. For each of the 15 gene sets, 10 bootstrap replicates were generated by sampling
610 genes randomly in each category. Genes in each replicate were concatenated and the Dfem
611 program run with the GammaExpo distribution of fitness effect (Galtier 2016). For each
612 recombination category, the mean estimates of α and ω_a, as well as the standard error over
613 the 10 replicates, were computed and plotted (Figure 2). Kendall correlation tests were
614 performed on the 15 times 10 estimates in order to correlate both α and ω_a with ρ.
615
616
617 **Model of codon sequence evolution and candidate genes under positive diversifying**
618 **selection**
619 We used all filtered CDS alignments to reconstruct genealogies for the individual genes using
620 PhyML (model HKY85) (Guindon and Gascuel 2003). To investigate patterns of selection and
621 infer the role of positive selection on adaptive gene evolution, the program CODEML from the
622 PAML package was used (Yang 2007) with the filtered multiple CDS alignments and the
623 corresponding phylogenetic trees as inputs. CODEML allows inference of selection and
624 evolutionary rates by calculating the parameter ω, the ratio of non-synonymous to
625 synonymous rates (dN/dS) for each gene. More specifically, we compared site models that
626 allow ω to vary among codons in the protein (Nielsen and Yang 1998). The models used in this
627 study include the nearly neutral (M1a), positive selection (M2a), beta (M7), and beta&ω (M8)
628 models. A likelihood ratio test (LRT) was used to compare the fit of null models and
629 alternative models, and the significance of the LRT statistic was determined using a χ^2
630 distribution. The first LRT tests for the occurrence of sites under positive selection by
631 comparing the M1a and M2a models. In the model M1a sites can be under purifying selection
632 ($0 < \omega < 1$) and evolve by neutral evolution ($\omega = 0$) while the M2a model allows for some sites
633 to be under positive selection ($\omega > 1$). The second LRT compares the M7 and M8 models,
634 where in M7 a discretized beta distribution for ω (limited to the interval [0,1]) over all sites is
635 assumed, while the model M8 allows sites with $\omega > 1$. By allowing for a wider range of
636 strength of purifying selection, the M7 and M8 models are more biologically realistic. The M7-
637 M8 LRT, however, was shown to display an increased false discovery rate compared to the
638 M1a-M2a comparison (Anisimova et al. 2001). For these two tests, p-values were obtained

639 with a χ^2 test (df = 2). We corrected for multiple testing and a false discovery rate of 1% was
640 used for the detection of genes under positive selection (Benjamini and Hochberg 1995). Only
641 genes significant for both tests were considered as genes evolving under positive selection.
642 To address divergent adaptation, we compared gene evolution among four closely related
643 *Zymoseptoria* species. In a previous study we defined the core proteome of *Z. tritici*, *Z.*
644 *ardabiliae*, *Z. brevis* and *Z. pseudotritici* comprising 7,786 orthologous genes (Grandaubert et
645 al. 2015). We generated alignments of the corresponding coding sequences using the MACSE
646 sequence aligner (Ranwez et al. 2011) and used CODEML with a branch model that allows ω
647 to vary among branches of the phylogeny (Nielsen and Yang 1998). As input we applied a non-
648 rooted tree of the four *Zymoseptoria* species as published in (Stukenbrock et al. 2012). Branch
649 lengths were re-estimated for each gene by CODEML (Yang 2007).
650
651 **Functional enrichment analysis and gene cluster analysis**
652 PFAM domains were extracted from Interproscan results from (Grandaubert et al. 2015). Only
653 domain hits with e-values lower than 1e-5 were considered resulting in 10,026 domains
654 present in 7,343 genes. Enrichment tests were performed based on contingency tables,
655 counting the number of genes containing the domain and the number of genes which do not
656 contain it, for both the complete proteome and a given set of candidates to test. A χ^2 test was
657 performed to assess significance.
658 To analyse the distribution of genes under positive selection, we considered two genes
659 separated by less than 5,000 bp to be clustered and assessed the probability of such clusters
660 under a random distribution of genes along the chromosomes. To do so on a genome-wide
661 scale, we calculated the probability to obtain clusters encompassing from two to ten genes
662 under positive selection when these genes are randomly distributed across all gene
663 coordinates. Based on 10,000 random permutations, it appeared that only clusters containing
664 more than three genes were significant at the 5% level.
665
666 **Genome-wide analysis of selection patterns**
667 We compared the selection regime of each gene to several variables. Local recombination
668 rate was computed using the fine-scale recombination map of Stukenbrock and Dutheil
669 (Stukenbrock and Dutheil 2017) as the average in the 500 bp 5' and 3' flanking regions. Local
670 gene density was computed as the proportion of coding sites in 10 kb upstream and
671 downstream regions. Similarly, the density of TEs was computed as the proportion of sites
672 annotated as TE in a windows starting 10 kb upstream the gene and ending 10 kb

673 downstream the gene, allowing for TE located within the gene itself. GC content at third
674 codon position (GC3) was also computed. Expression levels were calculated as in (Kellner et
675 al. 2014). Mean expression level was computed as the maximum value observed for the gene
676 in axenic culture or plant infection, each averaged over three biological replicates. Genes
677 encoding secreted proteins were predicted by the program SignalP (Petersen et al. 2011).
678 Genes with no homolog in any other species were considered as "orphan" genes and genes
679 located on accessory chromosomes were considered as "dispensable".
680 A linear model was defined with the strength of purifying selection (ω0) as a response
681 variable, together with all individual explanatory variables. We further included an interaction
682 of gene density and recombination rate, as expected under a background selection model and
683 an interaction between protein length given that orphan genes are reported to be
684 significantly shorter than other genes (Liao et al. 2006; Cai and Petrov 2010). Genes with ω0 <
685 0.001 or ω0 > 999.0 were discarded for further analyses as these estimates most likely
686 indicate model fitting issues, possibly due to lack of data. A BoxCox transform was used to
687 normalize the model residues. After normalization, a significant departure to normality was
688 still observed (Shapiro-Wilk normality test, p-value = 2.315), but no significant
689 heteroscedasticity (Harrison-McCabe test, p-value = 0.102), and model residues were found
690 to be independent (Box-Ljung test, p-value = 0.9023). We used the rms package in R to
691 compute robust regression estimates (Harrell 2015).
692 We selected the 10% *Z. tritici* most expressed genes and computed the relative synonymous
693 codon usage of every codons (Sharp et al. 1986). Analyses were conducted using the 'uco'
694 function of the seqinr package for R (R Core Team 2013).
695
696 **Simulation with recombination**
697 We used the MS program in order to simulated ancestral recombination graphs (ARGs) for
698 thirteen individuals and a given recombination rate (Hudson 2002). The resulting ARGs were
699 used to simulate codon alignments using the bppSeqGen program (Dutheil and Boussau
700 2008). Codon sites were simulated with dN/dS values of 0.1, 0.3, 0.6, 1.0 or 3.0, with equal
701 probability. In this set-up, recombination was effectively allowed only between codons and
702 not within codons. The resulting alignments were given as input to the PhyML program
703 (Guindon et al. 2005) with a HKY85 substitution model and a discretized gamma distribution
704 of site-specific rates with 4 classes, as used for inference of gene data sets in this study. M1a
705 and M2a models were then fitted using the estimated tree as input with CODEML.
706 Recombination rate of ρ = 0.0, 0.001, 0.01, 0.05 and 0.1, for gene lengths of 100, 500, 1,000

707 and 2,000 codons with 30 independent replicates for each combination of parameters.
708 CODEML output was parsed using BioPython (Cock et al. 2009).
709
710 **Substitution mapping**
711 In order to assess the putative occurrence of RIP mutations among mutations observed in a
712 given gene, we inferred the amount of CpA → TpA and GpT → ApT dinucleotide substitutions
713 using probabilistic substitution mapping with a dinucleotide substitution model previously
714 fitted to the data using maximum likelihood (Dutheil and Boussau 2008). Given a substitution
715 model and a phylogenetic tree, the probabilistic substitution mapping procedure enabled us
716 to count each type of substitution for a site of an alignment and each branch of a phylogeny,
717 while accounting for the uncertainty of the ancestral sequences at the inner nodes of the
718 phylogeny.
719

720 # Acknowledgements

721 The authors thank Nicolas Galtier and Thomas Bataillon for helpful discussion. The study was
722 funded by a Max Planck fellowship and a personal grant from the State of Schleswig-Holstein,
723 Germany both to EHS.
724
725

726 # References

727 Van den Ackerveken GFJM, Vossen P, De Wit J. 1993. The AVR9 race-specific elicitor of
728 Cladosporium fulvum is processed by endogenous and plant proteases. Plant Physiol.
729 103:91–96.
730 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
731 unrelated individuals. Genome Res. 19:1655–1664.
732 Ali S, Gladieux P, Leconte M, Gautier A, Justesen AF, Hovmøller MS, Enjalbert J, de Vallavieille-
733 Pope C. 2014. Origin, Migration Routes and Worldwide Population Genetic Structure of
734 the Wheat Yellow Rust Pathogen *Puccinia striiformis* f.sp. *tritici*. PLoS Pathog
735 10:e1003903.
736 André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M. 2014. CAZyme discovery
737 and design for sweet dreams. Curr. Opin. Chem. Biol. 19:17–24.
738 Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in
739 detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–1592.
740 BANKE S, MCDONALD BA. 2005. Migration patterns among global populations of the

741 pathogenic fungus Mycosphaerella graminicola. Mol. Ecol. 14:1881–1896.
742 Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful
743 approach to multiple testing. J. R. Stat. Soc. Ser. B:289–300.
744 Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the
745 effects of recombination on genome evolution. Mol Biol Evol 19.
746 Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K,
747 Clawson H, Green ED, et al. 2004. Aligning Multiple Genomic Sequences With the
748 Threaded Blockset Aligner. Genome Res. 14:708–715.
749 Brown JKM, Hovmøller MS. 2002. Aerial dispersal of pathogens on the global and continental
750 scales and its impact on plant disease. Science 297:537–541.
751 Brunner PC, Torriani SFF, Croll D, Stukenbrock EH, McDonald BA. 2013. Coevolution and life
752 cycle specialization of plant cell wall degrading enzymes in a hemibiotrophic pathogen.
753 Mol. Biol. Evol. 30:1337–1347.
754 Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in
755 primate lineage-specific genes. Genome Biol. Evol. 2:393–409.
756 Cambareri EB, Jensen BC, Schabtach E, Selker EU. 1989. Repeat-induced GC to AT mutations
757 in Neurospora. Science (80-.). 244:1571–1575.
758 Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between
759 recombination rate and patterns of molecular evolution and variation in Drosophila
760 melanogaster. Mol. Biol. Evol. 31:1010–1028.
761 Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on
762 neutral molecular variation. Genetics 134:1289–1303.
763 Charlesworth D. 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome
764 Regions. PLOS Genet. 2:e64.
765 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
766 Wilczynski B. 2009. Biopython: freely available Python tools for computational molecular
767 biology and bioinformatics. Bioinformatics 25:1422–1423.
768 Criscuolo A, Berry V, Douzery EJP, Gascuel O. 2006. SDM: a fast distance-based approach for
769 (super) tree building in phylogenomics. Syst. Biol. 55:740–755.
770 Croll D, Lendenmann MH, Stewart E, McDonald BA. 2015. The Impact of Recombination
771 Hotspots on Genome Evolution of a Fungal Plant Pathogen. Genetics 201:1213–1228.
772 Croll D, Zala M, McDonald BA. 2013. Breakage-fusion-bridge cycles and large insertions
773 contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. PLoS
774 Genet 9:e1003567.

775 Daverdin G, Rouxel T, Gout L, Aubertot J-N, Fudal I, Meyer M, Parlange F, Carpezat J,
776 Balesdent M-H. 2012. Genome Structure and Reproductive Behaviour Influence the
777 Evolutionary Potential of a Fungal Phytopathogen. PLOS Pathog. 8:e1003020.
778 Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast
779 protein evolution. Mol. Biol. Evol. 23:327–337.
780 Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev.
781 12:640–649.
782 Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon
783 usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. 96:4482–
784 4487.
785 Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++
786 suite of libraries and programs. BMC Evol. Biol. 8:1.
787 Dutheil JY, Gaillard S, Stukenbrock EH. 2014. MafFilter: a highly flexible and extensible
788 multiple genome alignment files processor. BMC Genomics 15:53.
789 Dutheil JY, Mannhaupt G, Schweizer G, Sieber CMK, Münsterkötter M, Güldener U, Schirawski
790 J, Kahmann R. 2016. A tale of genome compartmentalization: the evolution of virulence
791 clusters in smut fungi. Genome Biol. Evol. 8:681–704.
792 Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms.
793 Trends Ecol. Evol. 29:51–63.
794 Ellegren H, Galtier N. 2016. Determinants of genetic diversity. Nat. Rev. Genet.
795 Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol. Evol. 21:569–575.
796 Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat
797 Rev Genet 8:610–618.
798 Galtier N. 2016. Adaptive protein evolution in animals and the effective population size
799 hypothesis. PLoS Genet 12:e1005774.
800 Goodwin SB, M'barek S Ben, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, Foster AJ, Van der
801 Lee T a J, Grimwood J, Aerts A, et al. 2011. Finished genome of the fungal wheat
802 pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome
803 plasticity, and stealth pathogenesis. PLoS Genet. 7:e1002070.
804 Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov M V, Filatov DA, Eyre-
805 Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in
806 many plant species. Mol. Biol. Evol. 27:1822–1832.
807 Grandaubert J, Bhattacharyya A, Stukenbrock EH. 2015. RNA-seq Based Gene Annotation and
808 Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria

809 Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements.
810 G3 Genes|Genomes|Genetics 5(7):1323-33
811 Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D,
812 Pouyet F, Cahais V. 2013. Bio++: efficient extensible libraries and tools for computational
813 molecular evolution. Mol. Biol. Evol. 30:1745–1750.
814 Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large
815 phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
816 Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online—a web server for fast
817 maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33:W557–W559.
818 Harrell F. 2015. Regression modeling strategies: with applications to linear models, logistic
819 and ordinal regression, and survival analysis. Springer
820 Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet. Res.
821 (Camb). 8:269–294.
822 Hood ME, Katawczik M, Giraud T. 2005. Repeat-induced point mutation and the population
823 structure of transposable elements in Microbotryum violaceum. Genetics 170:1081–
824 1089.
825 Howlett BJ. 2006. Secondary metabolite toxins and nutrition of plant pathogenic fungi. Curr.
826 Opin. Plant Biol. 9:371–375.
827 Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic
828 variation. Bioinformatics 18:337–338.
829 Idnurm A, Howlett BJ. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-
830 induced point mutations can occur in the Dothideomycete Leptosphaeria maculans.
831 Fungal Genet. Biol. 39:31–37.
832 de Jonge R, van Esse HP, Kombrink A, Shinya T, Desaki Y, Bours R, van der Krol S, Shibuya N,
833 Joosten MHAJ, Thomma BPHJ. 2010. Conserved fungal LysM effector Ecp6 prevents
834 chitin-triggered immunity in plants. Science (80-.). 329:953–955.
835 de Jonge R, Peter van Esse H, Maruthachalam K, Bolton MD, Santhanam P, Saber MK, Zhang Z,
836 Usami T, Lievens B, Subbarao K V, et al. 2012. Tomato immune receptor Ve1 recognizes
837 effector of multiple fungal pathogens uncovered by genome and RNA sequencing. Proc.
838 Natl. Acad. Sci. U. S. A. 109:5110–5115.
839 Kellner R, Bhattacharyya A, Poppe S, Hsu TY, Brem RB, Stukenbrock EH. 2014. Expression
840 Profiling of the Wheat Pathogen Zymoseptoria tritici Reveals Genomic Patterns of
841 Transcription and Host-Specific Regulatory Programs. Genome Biol. Evol. 6:1353–1365.
842 Klosterman SJ, Subbarao K V, Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B,

843 Lee Y-H, Park J, et al. 2011. Comparative Genomics Yields Insights into Niche Adaptation
844 of Plant Vascular Wilt Pathogens. PLoS Pathog [Internet] 7:e1002137. Available from:
845 http://dx.doi.org/10.1371%2Fjournal.ppat.1002137
846 Kulkarni RD, Thon MR, Pan H, Dean RA. 2005. Novel G-protein-coupled receptor-like proteins
847 in the plant pathogenic fungus Magnaporthe grisea. Genome Biol. 6:R24.
848 Laurie JD, Ali S, Linning R, Mannhaupt G, Wong P, Guldener U, Munsterkotter M, Moore R,
849 Kahmann R, Bakkeren G, et al. 2012. Genome Comparison of Barley and Maize Smut
850 Fungi Reveals Targeted Loss of RNA Silencing Components and Species-Specific Presence
851 of Transposable Elements. Plant Cell 24:1733–1745.
852 Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast
853 distance-based phylogeny inference program. Mol. Biol. Evol. 32:2798–2800.
854 Liao B-Y, Scott NM, Zhang J. 2006. Impacts of Gene Essentiality, Expression Pattern, and Gene
855 Compactness on the Evolutionary Rate of Mammalian Proteins. Mol. Biol. Evol. 23:2072–
856 2080.
857 Linde CC, Zhan J, McDonald BA. 2002. Population Structure of Mycosphaerella graminicola:
858 From Lesions to Continents. Phytopathology 92:946–955.
859 Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-JM-J, Di Pietro A, Dufresne M,
860 Freitag M, Grabherr M, Henrissat B, et al. 2010. Comparative genomics reveals mobile
861 pathogenicity chromosomes in Fusarium. Nature 464:367–373.
862 Marais G, Charlesworth B. 2003. Genome Evolution: Recombination Speeds Up Adaptive
863 Evolution. Curr. Biol. 13:R68–R70.
864 Marshall R, Kombrink A, Motteram J, Loza-Reyes E, Lucas J, Hammond-Kosack KE, Thomma
865 BPHJ, Rudd JJ. 2011. Analysis of two in planta expressed LysM effector homologs from
866 the fungus Mycosphaerella graminicola reveals novel functional properties and varying
867 contributions to virulence on wheat. Plant Physiol. 156:756–769.
868 McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila.
869 Nature 351:652.
870 McDonald MC, McGinness L, Hane JK, Williams AH, Milgate A, Solomon PS. 2016. Utilizing
871 Gene Tree Variation to Identify Candidate Effector Genes in Zymoseptoria tritici. G3
872 Genes|Genomes|Genetics 6:779–791. t
873 Mehrabi R, Taga M, Kema GHJ. 2007. Electrophoretic and cytological karyotyping of the foliar
874 wheat pathogen Mycosphaerella graminicola reveals many chromosomes with a large
875 size range. Mycologia 99:868–876.
876 Möller M, Stukenbrock EH. 2017. Evolution and genome architecture in fungal plant

877 pathogens. Nat. Rev. Microbiol.
878 Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and
879 nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.
880 Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford university press
881 Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites
882 and applications to the HIV-1 envelope gene. Genetics 148:929–936.
883 Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on
884 background selection. Genet. Res. (Camb). 67:159–174.
885 O'Donnell K, Kistler HC, Tacke BK, Casper HH. 2000. Gene genealogies reveal global
886 phylogeographic structure and reproductive isolation among lineages of Fusarium
887 graminearum, the fungus causing wheat scab. Proc. Natl. Acad. Sci. 97:7905–7910.
888 Petersen TN, Brunak S, Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides
889 from transmembrane regions. Nat Methods 8.
890 Plissonneau C, Stürchler A, Croll D. 2016. The Evolution of Orphan Regions in Genomes of a
891 Fungal Pathogen of Wheat. MBio 7:e01231-16.
892 Poppe S, Dorsheimer L, Happel P, Stukenbrock EH. 2015a. Rapidly Evolving Genes Are Key
893 Players in Host Specialization and Virulence of the Fungal Wheat Pathogen Zymoseptoria
894 tritici (Mycosphaerella graminicola). PLoS Pathog 11:e1005055.
895 Poppe S, Dorsheimer L, Happel P, Stukenbrock EH. 2015b. Rapidly Evolving Genes Are Key
896 Players in Host Specialization and Virulence of the Fungal Wheat Pathogen Zymoseptoria
897 tritici (Mycosphaerella graminicola). PLoS Pathog. 11:e1005055.
898 Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, Zuccaro A, Reissmann S,
899 Kahmann R. 2015. Fungal effectors and plant susceptibility. Annu. Rev. Plant Biol.
900 66:513–545.
901 R Core Team. 2013. R: A language and environment for statistical computing.
902 Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, Jiang RHY, Zody MC,
903 Kunjeti SG, Donofrio NM, et al. 2010. Genome Evolution Following Host Jumps in the
904 Irish Potato Famine Pathogen Lineage. Science (80-.). 330:1540–1543.
905 Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding
906 SEquences accounting for frameshifts and stop codons. PLoS One 6:e22594.
907 Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret
908 L, Faivre N. 2014. Comparative population genomics in animals uncovers the
909 determinants of genetic diversity. Nature 515:261–263.
910 Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, Freitag M,

911 Stukenbrock EH. 2015. Histone modifications rather than the novel regional
912 centromeres of Zymoseptoria tritici distinguish core and accessory chromosomes.
913 Epigenetics Chromatin 8:1.
914 Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly
915 differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.
916 Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–
917 1024.
918 Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Loren van Themaat E Ver,
919 Brown JKM, Butcher SA, Gurr SJ, et al. 2010. Genome Expansion and Gene Loss in
920 Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism. Science (80-.).
921 330:1543–1546.
922 Stergiopoulos I, van den Burg HA, Ökmen B, Beenen HG, van Liere S, Kema GHJ, de Wit PJGM.
923 2010. Tomato Cf resistance proteins mediate recognition of cognate homologous
924 effectors from fungi pathogenic on dicots and monocots. Proc. Natl. Acad. Sci.
925 107:7610–7615.
926 Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, McDonald BA, Wang J,
927 Schierup MH. 2011. The making of a new pathogen: Insights from comparative
928 population genomics of the domesticated wheat pathogen Mycosphaerella graminicola
929 and its wild sister species. Genome Res. 21:2157–2166.
930 Stukenbrock EH, Dutheil JY. 2017. Comparison of fine-scale recombination maps in fungal
931 plant pathogens reveals dynamic recombination landscapes and intragenic hotspots.
932 bioRxiv
933 Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH. 2010. Whole-
934 Genome and Chromosome Evolution Associated with Host Adaptation and Speciation of
935 the Wheat Pathogen Mycosphaerella graminicola. PLoS Genet 6:e1001189.
936 Stukenbrock EH, Quaedvlieg W, Javan-Nikhah M, Zala M, Crous PW, McDonald BA. 2012.
937 Zymoseptoria ardabiliae and Z. pseudotritici, two progenitor species of the septoria
938 tritici leaf blotch fungus Z. tritici (synonym: Mycosphaerella graminicola). Mycologia
939 104:1397–1407.
940 Xue M, Yang J, Li Z, Hu S, Yao N, Dean RA, Zhao W, Shen M, Zhang H, Li C, et al. 2012.
941 Comparative Analysis of the Genomes of Two Field Isolates of the Rice Blast Fungus
942 Magnaporthe oryzae. PLOS Genet. 8:e1002869.
943 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–
944 1591.
945

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.

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

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

Effect	Coefficient	S.E.	P-value	
Intercept	1,5992	0,2323	<0.0001	***
Expression	0,0001	0,0001	0,109	NS
Recombination rate	-0,2419	0,0604	<0.0001	***
Protein size	-0,0004	0,0001	<0.0001	***
Gene density	-0,1098	0,091	0,228	NS
TE density	0,033	0,432	0,9391	NS
GC 3rd codon position	-5,6186	0,3441	<0.0001	***
Secreted	0,0796	0,0766	0,2985	NS
Dispensable	0,4291	0,2514	0,088	.
Orphan	0,9212	0,1834	<0.0001	***
Recombination rate * Gene density	-0,9378	0,3509	0,0076	**
Protein size * Orphan	-0,0012	0,0006	0,0444	*

Figures

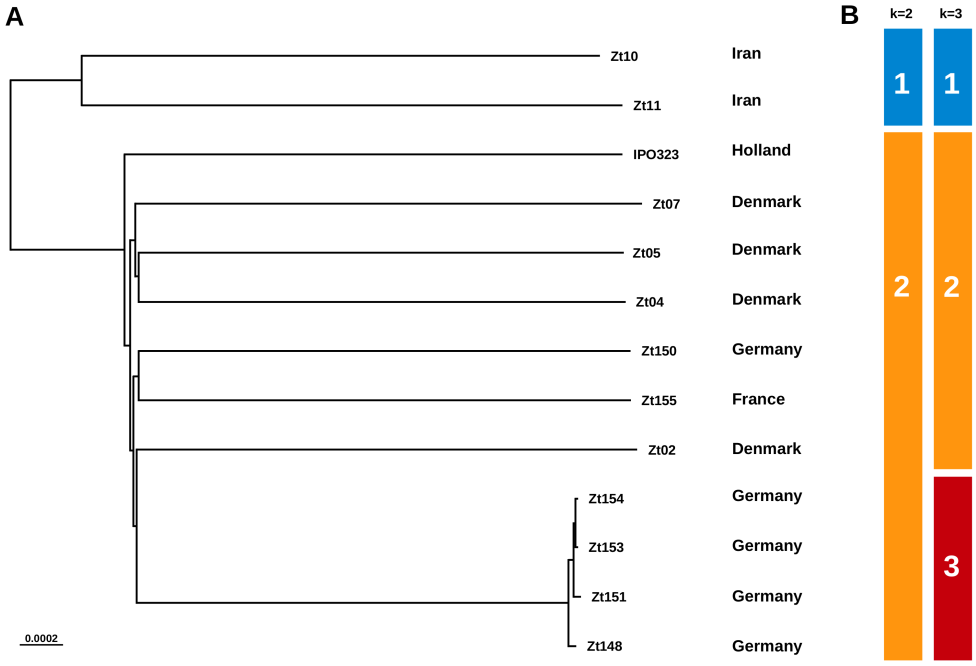

Figure 1: Population structure of the thirteen *Z. tritici* isolates. A) Consensus super tree of the thirteen isolates based on 1,850 genealogies estimated in 10 kb sliding windows along the multiple genome alignment. This tree suggests the grouping of the isolates into two populations originating from 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.

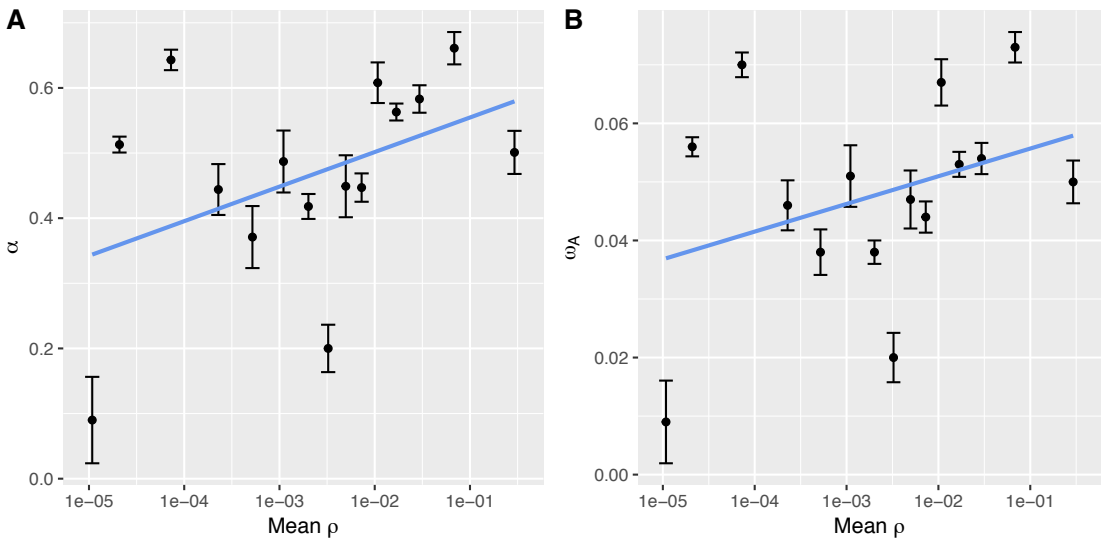

Figure 2: Estimates of A) the proportion of adaptive substitution α, and B) the rate of adaptive substitution, ω_a as a function of the population recombination rate (ρ). Each point and bars

represent the mean estimate and corresponding standard error for one recombination category over 10 bootstrap replicates. The blue lines show the fitted linear models.

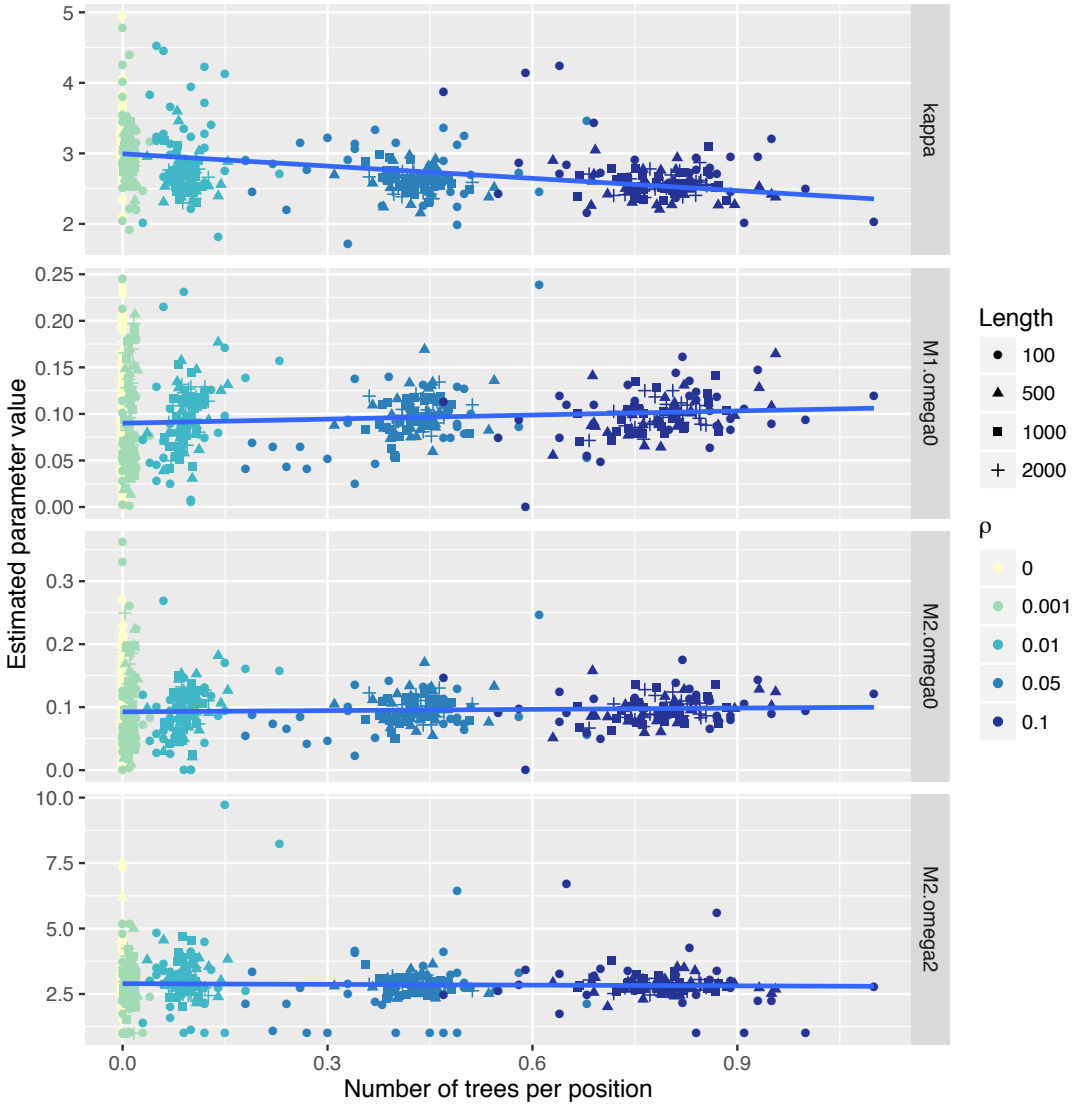

Figure 3: Impact of recombination on the estimation of the strength of purifying selection using phylogenetic models of codon sequence evolution. Parameters are the ratio of transitions over transversions (kappa), the ratio of the rates of synonymous over non-synonymous mutations for sites evolving under 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.

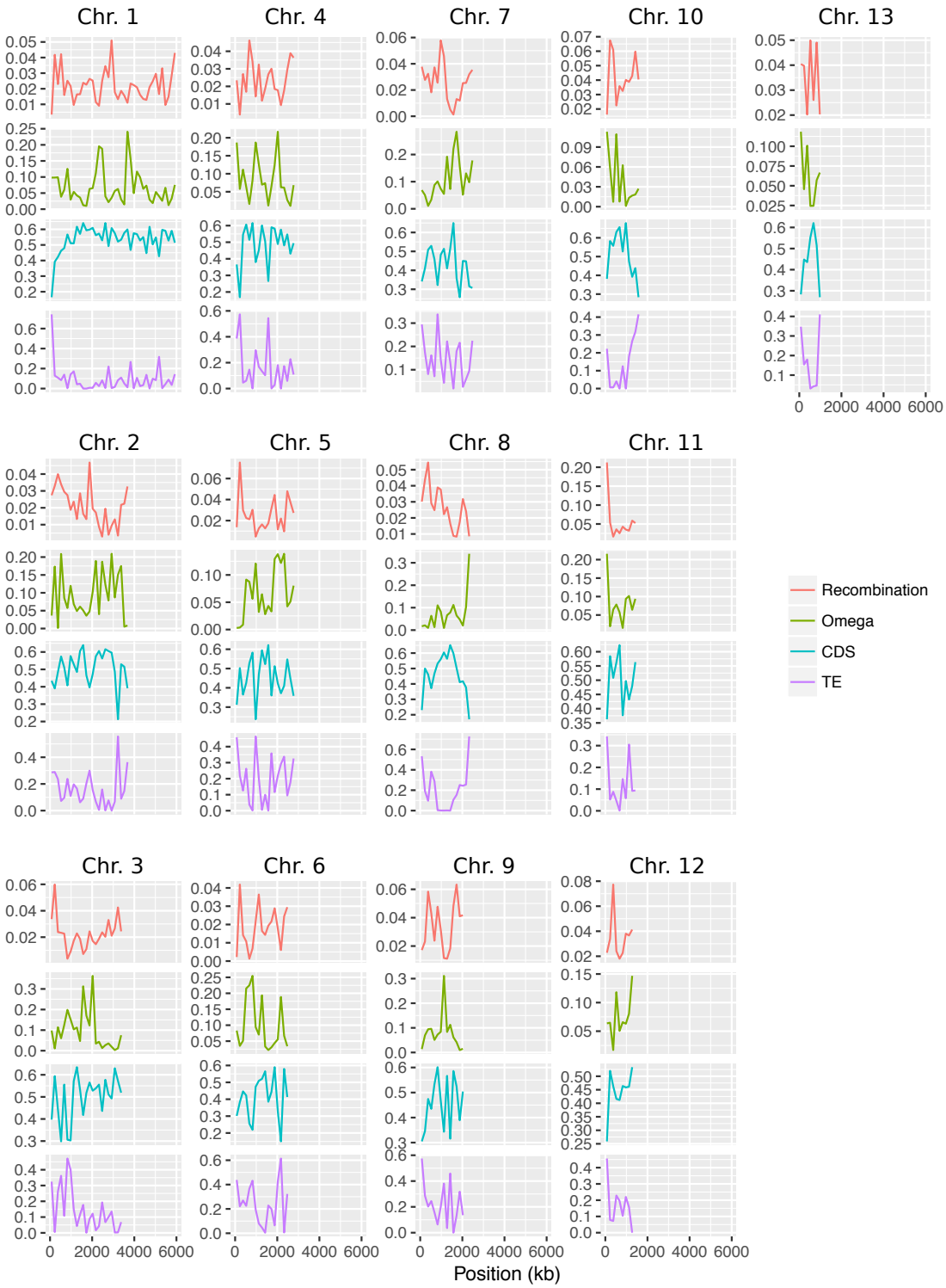

Figure 4: Patterns of selection along the genome of Z. *tritici*. Recombination, intensity of purifying selection (as estimated from the Pn/Ps ratio), density of coding sites (CDS) and density of transposable elements (TE) are plotted in windows of 150 kb along the thirteen essential chromosomes.

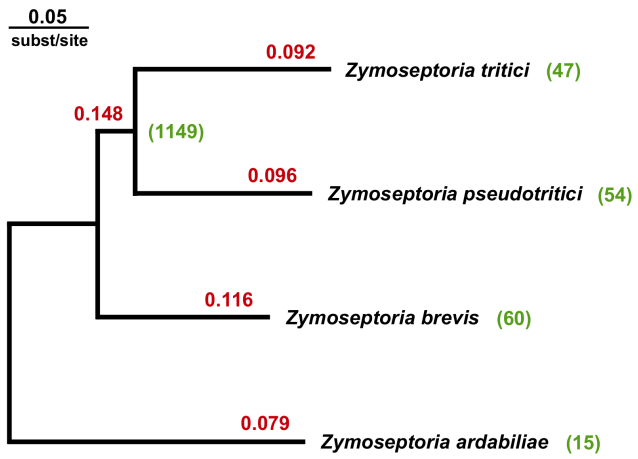

Figure 5: Branch-specific signatures of divergent adaptation in *Zymoseptoria* species. Phylogeny of four *Zymoseptoria* species based on integrating the mean dN / dS ratio (in red) and the number of positively selected genes (in green) estimated from the 7,786 genes shared by the species using the branch 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.