Title: The genomic determinants of adaptive evolution in a fungal pathogen

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)] UMR 5554 Institut des Sciences de l'Evolution, CNRS, IRD, EPHE, Université de Montpellier, Place E. Bataillon, 34095, Montpellier, 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

Runing Title: Adaptive evolution in a fungal plant pathogen

Abstract:

Background: 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*.

Results: 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. We report that the pN/pS ratio of sites under purifying selection negatively correlates with the local recombination rate, as expected under a scenario of background selection. Contrasting polymorphism and divergence with a closely related species, 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 of sequence evolution, we furthermore identify a set of 786 additional genes showing signatures of diversifying selection, which we find are preferentially located in region of high recombination rates.

Conclusions: We report a high rate of adaptive mutations in the genome of *Z. tritici*, in particular in genes involved in host-pathogen interactions. Furthermore, we identify a set of genes with signature of balancing / diversifying selection, which we find are related to secondary metabolism and host infection. Finally, we report a strong impact of recombination on the efficacy of purifying, adaptive and diversifying selection, suggesting pervasive effect of linked selection in the genome and thereby highlighting the importance of recombination in this pathogen species.

Background

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 [1]. 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 [2]. 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. [3–6]). Repetitive DNA may locally increase mutation rate and contribute to gene duplications and structural variation among alleles [7]. 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 [8]. 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 [9]. Further tests, building on pioneering work by McDonald and Kreitman [10], 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 directional 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 [11,12].

100 The McDonald-Kreitman table forms the basis for the quantification of the proportion of
101 adaptive substitutions in a gene fixed by positive selection defined as $\alpha = 1 - (pN / pS) / (dN /$
102 $dS) = 1 - (dS * pN) / (dN * pS)$ [13]. Additional modeling of the fitness effect of substitutions
103 allows α to be estimated in the presence of slightly deleterious or slightly advantageous
104 mutations that otherwise can bias the inference of positively selected substitutions [14].
105 These methods have been applied to a range of different species and have been used to show
106 that the rate of adaptive evolution varies largely among different taxa [11,12].
107 In pathogen genomes, increased rates of adaptation, as occurring by arms-race evolution
108 where new non-synonymous mutations repeatedly are selected for, will result in higher
109 dN/dS ratios. Indeed, evolutionary predictions based on dN/dS analyses and McDonald-
110 Kreitman tests have identified effectors and other pathogenicity-related genes in pathogen
111 genomes [15,16]. However, genes that evolve under a trench warfare scenario leave different
112 signatures in the genome sequence. This type of selection can also be referred to as positive
113 diversifying selection where a diversity of alleles at the selected locus is maintained by natural
114 selection. The genic signature of positive diversifying selection is an increased proportion of
115 non-synonymous polymorphisms [17]. Yet, in a functionally important gene, the majority of
116 non-synonymous sites will be evolving under selective constraints, and only a small number
117 will be allowed to change without negatively affecting protein function. Therefore, in most
118 cases, searching for signatures of positive diversifying selection simply by comparing the
119 average proportion of non-synonymous to synonymous polymorphisms in a gene will show
120 little statistical power [18]. 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) [18].
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 [19]. The accessory chromosomes comprise repeat rich,
129 heterochromatic DNA with a low gene content and a so far unknown functional relevance
130 [20,21]. We have previously shown that evolutionary rates on these chromosomes are
131 particularly high suggesting that genes on the accessory chromosomes in general evolve
132 under less selective constraints [22].

133 Previous studies based on comparative population genomic analyses of *Z. tritici* and two
134 closely related species, *Z. pseudotritici* and *Z. ardabiliae* used genome-wide estimates of non-
135 synonymous and synonymous divergence to identify past species-specific signatures of
136 selection in the wheat pathogen [16,22]. Functional characterization of some of these genes
137 revealed amino acid substitutions important for *in planta* development and asexual spore
138 formation in the wheat-adapted pathogen, and thereby confirmed the use of evolutionary
139 predictions to identify functionally relevant traits [23].
140 While past signatures of selection can be related to ecological specialization of *Zymoseptoria*
141 species, ongoing positive selection likely reflects host-pathogen arms race or trench warfare
142 evolution and recurrent adaptations to changing environmental conditions including fungicide
143 treatments and wheat cultivars. Population genomic studies of *Z. tritici* have revealed that the
144 pathogen has a high effective population size, which correlates positively with a strong
145 efficacy of natural selection, i.e. the ability to rapidly fix advantageous mutations and
146 eliminate non-advantageous [16]. Additional parameters that can influence rapid adaptation
147 are recombination and mutation rates, as well as genome architecture. Recombination rate
148 variation has been studied in *Z. tritici* by experimental mating and population genomic
149 sequencing. Independently, these approaches have confirmed exceptionally high
150 recombination rates (~60cM/Mb) ([24,25]). So far the implications of such high recombination
151 rates on gene evolution are not known.
152 We apply a population genomics approach to infer genome-wide positive and purifying
153 selection among thirteen isolates of *Z. tritici* collected from bread wheat in Europe and the
154 Middle East. Our analyses based on more than 1.4 million single nucleotide polymorphisms
155 (SNPs) and including 700,000 coding sites allow us to identify past and on-going signatures of
156 selection in the genome of *Z. tritici*. Notably, our analyses reveal a strong importance of
157 recombination in gene evolution and a particularly high rate of adaptive substitutions in
158 genes encoding putative effectors. These analyses underline the potential of rapid adaptation
159 of virulence related traits in this important agricultural pathogen.
160

161 Results and Discussion

162 Genetic structure of *Z. tritici* populations

163 We generated a population genomic dataset of thirteen *Z. tritici* isolates obtained from
164 different field populations in Europe and Iran (Table S1). Given the high extent of structural
165 variation in genomes of *Z. tritici* isolates, we assembled each genome *de novo* to capture
166 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. [26,27]). 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 [28] 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* [29,30], 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 [31].

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 [32–34]. Based on the multiple genome alignment, we here investigated presence-absence polymorphisms of chromosomes,

201 chromosome fragments and genes among the thirteen isolates. We focused on non-repetitive
202 DNA and show, in agreement with a previous study [35], that 96.4% of the core chromosomes
203 (chromosome 1 to 13) are shared among all isolates (Figure S1). However, as expected, the
204 mapping of assembled contigs aligning to the accessory chromosomes varied considerably
205 between isolates (Figure S1).
206 Overall, 9,499 of the 11,839 genes predicted in the reference genome of IPO323 are present
207 in all of the thirteen isolates. Of the remaining, 1,584 genes are located on core chromosomes
208 and 667 on accessory chromosomes. This set of genes comprises genes present at different
209 frequencies among the thirteen *Z. tritici* isolates including 89 genes present only in IPO323
210 and of which 77 locate on core chromosomes.
211 A previous genome comparison of two Swiss *Z. tritici* isolates revealed nearly 300 novel genes
212 not characterized in the reference genome [36]. We speculated that the different isolates of
213 *Z. tritici* also could contain accessory genes and chromosomes not present in the reference
214 IPO323. To test this we investigated the subset of contigs that were not aligned to the
215 reference genome. We found a total of 5.2 Mb in the twelve re-sequenced *Z. tritici* isolates
216 with no homology to sequences in the reference genome. These new *Z. tritici* sequences,
217 corresponding to accessory fragments, encompass between 602 kb and 995 kb of non-
218 repetitive DNA in each genomes. We applied an *ab initio* gene predictor to assess whether
219 these accessory sequences encode proteins and found evidence for 1,041 novel full-length
220 gene models encoding proteins of at least 30 amino acids. Of these 75% encode proteins
221 without known function consistent with a general pattern observed on the accessory
222 chromosomes with a high proportion of gene models without a predicted function [19].
223 Isolate-specific fragments were previously shown to encode effector candidate genes strongly
224 up-regulated in *Z. tritici* during host infection [36]. In other fungal pathogens such as
225 *Verticilium dahliae* and *Magnaporthe oryzae* isolate-specific genes were experimentally
226 shown to encode virulence determinants [37,38]. Based on our findings, we hypothesize that
227 the large number of isolate-specific genes in thirteen *Z. tritici* genomes could play a role in
228 host-pathogen interactions. Under this scenario, the observed variation in gene content could
229 reflect diversifying selection imposed by genetic traits in a variety of host genotypes.
230
231 **Measures of adaptive evolution in *Z. tritici***
232 We next aimed to obtain a quantitative assessment of adaptive mutations in the genome of *Z.*
233 *tritici*. To this end, we first estimated the non-synonymous and synonymous divergence dN
234 and dS using a genome alignment of *Z. tritici* and its sister species *Z. ardabiliae*. Furthermore,

235 we used the *Z. tritici* SNP data to compute the (unfolded) site frequency spectra (SFS) of
236 synonymous and non-synonymous sites using *Z. ardabiliae* as an outgroup. By contrasting
237 divergence and polymorphism data, we estimated the parameter α as a measure of the
238 proportion of adaptive non-synonymous substitutions. The SFS is however strongly affected
239 by demography and the presence of slightly deleterious mutations segregating at low
240 frequencies (see review by [14]). State-of-the-art statistical methods account for these effects
241 by modeling the distribution of fitness effects (DFE) of mutations [12,39]. We used the
242 program Dfem to compute α as well as ω_a, the rate of adaptive substitutions, using four
243 distinct DFE models accounting for mutations with both slightly deleterious and beneficial
244 effects [12]. We fitted and compared the DFE models (see Materials and Methods) and found
245 that the Gamma-Exponential model best fitted our data (Table 1) in agreement with studies
246 from animals [12]. This suggests the existence of slightly deleterious, as well as slightly
247 beneficial segregating mutations in the genome of *Z. tritici* (Table 1). The estimates provide an
248 α value of 44% as a genome average, and an ω_a value of 0.05. Both values are in the range of
249 what is observed for Mammals (with the exception of Primates) but considerably higher than
250 estimates from plants [12,39]. In candidate effector genes, however, the rate of mutations
251 fixed by selection is more than twice as high as the average of all genes (ω_a equal to 0.14
252 compared to the genome average of 0.05, Table 1), with 68% of substitutions in these genes
253 inferred to be adaptive. This average estimate is close to the highest values reported in
254 animals [12]. We further aimed at assessing the impact of recombination on the fixation of
255 adaptive mutations, by considering genes in distinct recombination environments. We report
256 a significant positive correlation between α and ρ (the population scaled recombination rate
257 inferred as ρ = 4Ne x r) (Kendall's tau = 0.25, p-value = 6.816e-6) and ω_a and ρ (Kendall's tau =
258 0.21, p-value = 1.95e-4, Figure 2), suggesting that a higher recombination rate favors the
259 fixation of adaptive mutations, as expected by a Hill-Robertson effect [40,41]. These results,
260 based on population genomic data of *Z. tritici*, represent the first quantitative assessment of
261 adaptive evolution in the genome of a fungal pathogen. The exceptionally high rate of
262 adaptation in effector genes likely reflects arms race evolution driven by the antagonistic
263 interaction of *Z. tritici* and its host.
264
265 **Strength of purifying selection along the genome of *Z. tritici***
266 Loci at different positions of the genome evolve at different speed. This is due to variation in
267 the effect of different evolutionary processes across the genome as well as variation in the
268 composition of the DNA sequence itself [42]. In sexual organisms, recombination is a strong

269 determinant of evolutionary rates, and differences in the frequency of recombination can
270 strongly impact the rate whereby adaptive mutations can be fixed and whereby (slightly)
271 deleterious mutations are removed [41]. Evolutionary processes at any given site in the
272 genome are strongly influenced by selection acting on linked sites, a process referred to as
273 linked selection [40] and background selection in the case of linkage with mutation under
274 purifying selection [40,43]. Here, we aimed to address the genome-wide strength of purifying
275 selection in *Z. tritici* in protein coding genes, using the ratio of non-synonymous to
276 synonymous polymorphisms (pN/pS ratio).
277 We computed pN and pS for each gene with two methods: (1) by inference of average
278 pairwise heterozigosity [44,45] and (2) using a maximum likelihood framework and models of
279 sequence evolution [18]. In this framework, two types of models were compared: a two-rate
280 model (M1a), with sites evolving under neutrality or purifying selection with a pN/pS ratio
281 equal or below one (neutral model with purifying selection) and a three-rate model (M2a)
282 allowing for some sites to evolve under positive selection with a pN/pS ratio above one
283 (positive selection model) [46]. The best fitting model according to a likelihood ratio test
284 (False Discovery Rate < 0.01) was kept for further analyses. Such model-based inference
285 specifically accounts for heterogeneous selective pressures within genes, allowing inference
286 of pN/pS ratios for sites under purifying selection and for sites under positive selection, if any.
287 The average heterozygosity computation assumes an infinite site model and does not allow
288 for recurrent mutations. The model-based approach, while accounting for recurrent
289 mutations and more accurate mutational models, assumes an underlying tree structure and
290 no recombination in the region under study [47]. As *Z. tritici* exhibits both high mutation rates
291 and intragenic recombination, it is a priori unclear which method provides the best estimates
292 of pN/pS ratios and purifying selection. To assess the sensitivity of model-based estimates in
293 the presence of recombination, we used coalescent simulations (See Methods). We find that
294 estimates of purifying selection based on the pN/pS ratio (denoted "ω0" in the M1a and M2a
295 models) are indeed inflated in the presence of high recombination rates (Figure S2). While the
296 effect is significant, it is of small size. Furthermore, it goes in the opposite direction of what is
297 expected under background selection, as selection should be more efficient in highly
298 recombining regions and lead to the removal of more deleterious mutations reflected by a
299 lower omega ω0 value. We therefore conclude that the model-based inference of purifying
300 selection is conservative and we used these estimates in the further analyses. We estimated
301 the strength of purifying selection for a gene as the ω0 parameter of the best fitting model

302 between M1a and M2a. This procedure ensures that we only measure the strength of
303 purifying selection, even in the presence of positions evolving under positive selection.
304 To investigate which genome parameters determine the strength of purifying selection in *Z.*
305 *tritici*, we compared ω0 for each gene to 1) mean gene expression, 2) population
306 recombination rate in flanking non-coding regions, 3) gene length, 4) local gene density and 5)
307 GC content at third codon positions. We further considered whether the gene 6) is predicted
308 to be secreted, 7) is located on an accessory chromosome and 8) is an orphan gene (Figures 3
309 and 4). We fitted a linear model with ω0 as a response variable and the above-mentioned
310 observed statistics as explanatory variables. We identify several variables that significantly
311 explain the strength of purifying selection (Table 2): the length of the coding protein (number
312 of codons) has a highly significant negative effect, meaning that evolution of larger genes is
313 more constrained. Other studies that document an impact of gene length on evolutionary
314 rates have likewise shown that small genes evolve faster than long genes [48]. We also
315 observe that orphan genes, which in general are shorter than core genes (mean 325 amino
316 acids vs 516 amino acids on average, Wilcoxon test with p-value < 2.2e-16), are less
317 conserved, even when protein length is accounted for (Table 2) [48]. Many orphan genes in *Z.*
318 *tritici* have likely evolved de novo and may not encode proteins of essential functions. Our
319 observation of higher ω0 in orphan genes is consistent with findings in other organisms, e.g.
320 in primates where younger lineages specific genes evolve under less selective constraints than
321 older "common" primate genes [49].
322 We used *Z. tritici* gene expression data from early host colonization (four days after spore
323 inoculation on leaves of seedlings of a susceptible wheat host) and *in vitro* growth to assess
324 the correlation of transcription level and gene evolution [50]. Studies in other organisms
325 including yeast and bacteria have documented a strong impact of expression levels on gene
326 evolution whereby highly expressed genes are more conserved [48,51]. Consistent with these
327 observations we also find a significant negative correlation between ω0 and gene expression
328 (Figure 4 and Table 2) showing that highly expressed genes are more conserved in *Z. tritici*.
329 We furthermore find a highly significant negative effect of recombination rate on ω0,
330 implying that purifying selection is more efficient in regions with high recombination rate
331 (Figures 4 and 5). This result is in agreement with a model of background selection, where
332 purifying selection at linked loci reduces the local effective population size, therefore
333 reducing the efficacy of selection and allowing slightly deleterious mutations to spread more
334 frequently than at unlinked loci [43,52]. This finding supports a central role of recombination
335 in the removal of non-adaptive mutations in the genome of *Z. tritici* consistent with patterns

336 described in *Drosophila melanogaster* where rates of evolution were compared in regions of
337 high and low recombination [53].
338 Lastly, we find that GC content is significantly negatively correlated with ω0, that is, genes
339 with higher GC tend to be under stronger purifying selection. This may be explained by an
340 indirect effect of recombination as recombination rate and the GC content at third codon
341 positions (GC3) are positively correlated (Figure 4, Kendall's tau = 0.097, p-value < 2.2e16). A
342 similar correlation of recombination and GC3 is found in other organisms [54]. In
343 *Saccharomyces cerevisiae* this correlation has been explained by the impact of biased gene
344 conversion on sequences evolution [55]. However, a thorough search for signatures of GC-
345 biased gene conversion did not find any pervasive effect of this phenomenon in *Z. tritici* [25].
346 An alternative explanation is selection for codon usage. In *Z. tritici,* as in several other
347 organisms, preferred codons often contain a cytosine or a guanine at their third position
348 (Figure S3) [56]. As highly expressed genes are more conserved and enriched in preferred
349 codons, the relationship between ω0 and GC might be a by-product of the correlation with
350 gene expression.
351
352 **Detection of on-going diversifying selection in *Z. tritici***
353 Next we aimed at detecting sites in the *Z. tritici* genome showing signatures of positive
354 diversifying selection. We fitted models of codon sequence evolution as implemented in the
355 CodeML program of the PAML package to detect genes with significant signatures of positive
356 diversifying selection using likelihood ratio tests [46]. Two models are typically compared: the
357 "null" model M1a, where sites evolve only under neutral or purifying selection (ω <= 1), and
358 an alternative model M2a where sites in coding sequence also can evolve with a ratio of non-
359 synonymous over synonymous mutations above one (ω > 1). When applied to population
360 data, sites with ω > 1 can reflect positive diversifying selection [47].
361 We fitted codon models for genes present in at least three isolates (83% of genes located on
362 core chromosomes and 31% of genes on the accessory chromosomes) (Table S3). After
363 correcting for multiple testing, we identified a final set of 786 genes (including 24 on the
364 accessory chromosomes) having sites evolving under positive diversifying selection (False
365 Discovery Rate < 0.01) (Table S3). In total, our analyses recognized 4,686 individual codons
366 evolving under positive selection in the *Z. tritici* genome (0,001% of 4,640,090 analyzed
367 codons).
368 In several pathogen genomes, rapidly evolving genes are found clustered in particular
369 genomic environments often associated with repetitive sequences (e.g [7,15]). To address if

370 the same pattern is found in *Z. tritici*, we assessed the spatial distribution of genes under
371 positive diversifying selection along chromosomes (see Materials and Methods). Of the 786
372 positively selected genes, 240 are located within a distance of 5 kb from each other and
373 thereby form 108 clusters containing two to four genes with signatures of positive selection.
374 At the genome scale, clusters containing two or three genes do not show a significant pattern
375 as the same pattern can be obtained by randomly distributing the positively selected genes
376 across the genome. However, there are two significant clusters (p-value = 1.8×10^{-3})
377 containing four and eight genes, respectively. The clusters comprising four genes is located in
378 a 24kb region of chromosome 5, and the cluster comprising eight genes in a 31 kb region of
379 chromosome 9. For genes in both clusters no functional relevance can be assigned. Aside
380 these two clusters, there is no evidence that positively selected genes are physically
381 associated with each other in *Z. tritici*, nor that they are significantly enriched or depleted in
382 repeat rich regions.
383 The molecular battlefield between plants and their pathogens involves proteins secreted from
384 both host and pathogen cells [2]. Pathogenicity related genes therefore often encode signal
385 peptides targeting the translated protein for secretion. As shown above, adaptive
386 substitutions are fixed at a much higher rate in candidate effectors than in non-effector
387 genes. This observation is in agreement with an antagonistic co-evolution of genes involved in
388 host-pathogen protein-protein interactions. We also asked whether positive diversifying
389 selection, consistent with a trench-warfare scenario of co-evolution, drives effector evolution
390 in *Z. tritici*. To this end we assessed the proportion of effector candidates among the genes
391 predicted to be under diversifying selection, but found no evidence for a particular
392 enrichment of these genes (X^2 test, p-value = 0.171).
393 To jointly analyse the potential genomic variables affecting the distribution of genes under
394 positive diversifying selections, we fitted a binary logistic regression where the occurrence of
395 positive selection, as predicted by the likelihood ratio test on codon models, is used as
396 response variable. Several explanatory variables were included (Table 3), and three of them
397 appeared to have a highly significant effect. First, we report that highly expressed genes are
398 less likely to be under positive selection, which we hypothesise is due to a higher level of
399 selective constraint. Next, we find that genes in regions with higher recombination rates are
400 more likely to be under positive diversifying selection, which we interpret as evidence for a
401 Hill-Robertson effect [41,53]. Lastly, the longer the gene the more likely it is to be under
402 positive selection. We speculate however that this effect reflect a higher statistical power is
403 large genes.

404 We further extended our analyses of the genes predicted to be under positive diversifying
405 selection by characterizing known proteins domains. From the 786 candidate genes, 602 of
406 the encoded proteins (76.5%) have an *in silico* attributed function or harbour known protein
407 domains. We conducted a PFAM domain enrichment analyses and identified 21 significantly
408 enriched domains (FDR <= 0.05) (Table S4). These domains can be grouped into different
409 categories based on their associated molecular function and include carbohydrate-active
410 enzymes (CAZymes), polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS),
411 and cellular transporters (Table S4). Several of these categories are relevant for the pathogen
412 to interact with its host. For example, CAZymes are proteins involved in the break down of
413 glycosidic bonds contained in plant cell walls [57], and PKS and NRPS are multimodular
414 enzymes involved in the biosynthesis of secondary metabolites of which many also have been
415 shown to be involved in virulence of plant pathogenic fungi (reviewed in [58]).
416 Among the positively selected *Z. tritici* genes, we also searched for homologs of known
417 virulence factors described in other plant pathogens. The PFAM domain PF14856 corresponds
418 to the mature part of a virulence factor Ecp2 described in the tomato fungal pathogen
419 *Cladosporium fulvum* [59]. Ecp2 has been described in several other plant pathogens [60] and
420 has three homologs in *Z. tritici*. We find that two of these homologs comprise sites under
421 positive selection supporting a virulence related role of this gene also in this wheat pathogen.
422
423 **Transposable elements and the origin of new adaptive mutation**
424 In some filamentous Ascomycetes, Repeat Induced Point (RIP) mutations are a genome
425 defence mechanism that targets repetitive DNA. While experimental evidence for RIP so far
426 only has been documented in a few species [61,62], sequence signatures of RIP are reported in
427 several other Ascomycete and Basidiomycete species [63,64]. During meiosis RIP generates
428 single nucleotide changes mutating cytosine bases to thymine, preferentially at CpA
429 dinucleotides [61]. The accelerated mutation rate in repetitive DNA or duplicated sequences
430 mediated by RIP can interfere with the evolution of non-repetitive DNA via "leakage" of RIP
431 mutations into neighboring genes [6]. To infer the potential impact of RIP-like mutations on
432 gene evolution in *Z. tritici* we assessed the proportion of CpA → TpA (and GpT → ApT for the
433 reverse strand) mutations in all dinucleotide substitutions using a substitution mapping
434 approach for genes located at a maximum distance of 500, 1,000, 1,500 and 2,000 bp from a
435 TE. We find a significant enrichment of RIP-associated dinucleotide substitutions in genes
436 located close to TEs suggesting "leakage" of RIP mutations (Table S5). The proportions of RIP
437 mutations decrease with the distance to TEs supporting that the CpA → TpA mutations indeed

438 originate from RIP. However, we find no enrichment of positively selected sites among the RIP-
439 associated mutations, suggesting that in *Z. tritici*, RIP is not a major driver of adaptive
440 evolution.
441
442 **Signatures of past selection in *Z. tritici* and related species**
443 In a previous study the evolutionary history of *Z. tritici* was inferred using a whole genome
444 coalescence analyses [16]. We showed that divergence of *Z. tritici* and its sister species *Z.*
445 *pseudotritici* occurred recently and likely coincides with the onset of wheat domestication
446 and thereby specialization of *Z. tritici* to a new host. We hypothesize that genes important for
447 the colonization of distinct hosts have been under selection during the divergence of
448 *Zymoseptoria* species. To infer signatures of past selection we applied the branch model
449 implemented in the program package PAML to estimate the branch-specific dN/dS ratio for
450 core *Zymoseptoria* genes (present in four species *Z. tritici*, *Z. pseudotritici*, *Z. ardabiliae* and *Z.*
451 *brevis*[20] [46]. The branch-specific dN/dS ratios reflect the proportion of non-synonymous to
452 synonymous substitutions accumulated in each branch of the *Zymoseptoria* phylogeny. Our
453 analyses identified 47 genes with a dN/dS ratio > 1 on the *Z. tritici* branch indicative of an
454 increased non-synonymous divergence, 54 genes in *Z. pseudotritici*, 60 genes in *Z. brevis* and
455 15 genes in *Z. ardabiliae* (Figure 5). Based on their putative function in host-pathogen
456 interactions, we hypothesized that some positively selected genes encode secreted proteins
457 and putative effectors. Indeed we find a significant enrichment of putative effectors among
458 the positively selected genes on the *Z. tritici* branch (10 genes encoding effector candidates,
459 χ^2 test, p = 5.272 x 10^{-05}), and the *Z. pseudotritici* branch (9 genes encoding effector
460 candidates, χ^2 test, p = 0.00315). For *Z. ardabiliae* and *Z. brevis* the branch-specific positively
461 selected genes were however not enriched in effector candidates (in *Z. ardabiliae* 1 gene
462 encoding an effector candidate, χ^2 test, p=0.768, and in *Z. brevis* 8 genes encoding effector
463 candidates, χ^2 test, p=0.1165). We next performed a PFAM domain enrichment analysis for
464 the positively selected genes with a predicted function to address if some functional domains
465 are significantly enriched in this set of genes (FDR < 0.05) (Table S6). The majority of genes
466 encode proteins of unknown function, however among the functionally characterized
467 proteins we find a gene encoding a regulator of chromosome condensation that was
468 previously described to be functionally relevant for virulence in *Z. tritici* (Poppe et al. 2015).
469 Our analyses reveal an enrichment of genes encoding proteinases, one gene encoding Lysin
470 motifs (LysM) already described as an effector in *Z. tritici* and other fungal pathogens [66,67],
471 and one gene encoding a CFEM domain. Cystein-rich CFEM domains have been described in

472 G-protein-coupled receptors in the rice blast pathogen *Magnaporthe oryzae* playing an
473 important role in pathogenicity [68]. These genes provide interesting candidates for further
474 studies of molecular determinants of host specificity in *Z. tritici*.
475
476

477 Conclusions

478 The analysis of the assembled genomes of thirteen isolates of the fungus *Z. tritici* revealed a
479 high nucleotide diversity, including structural variability and gene presence-absence
480 polymorphisms. Based on models of codon sequence polymorphism and divergence, the
481 analyses presented here provide evidence for signatures of both positive directional and
482 diversifying selection in protein coding genes, as expected under arms race and trench
483 warfare scenarios, respectively. The rate of adaptive substitution in the plant pathogen was
484 found to be similar to estimates in animal species and considerably faster than corresponding
485 estimates in plants. Notably, rates of evolution in genes encoding effector proteins were
486 found to be more than twice as fast as the genome average. Furthermore, our results suggest
487 widespread occurrence of linked selection (both Hill-Robertson interference and background
488 selection), as both positively and negatively selected loci were found to impact the genetic
489 diversity at linked sites. These results highlight the fundamental role of recombination and
490 sexual reproduction in adaptive processes of rapidly evolving organisms.
491

492 Materials and Methods

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

494 In this study we used a geographical collection of thirteen field isolates of *Z. tritici* isolated
495 from infected leaves of bread wheat (*Triticum aestivum*) (Table S1). Genome data of three
496 isolates, including the reference isolate IPO323, were published previously [16,19]. For the
497 remaining ten isolates full genomes were sequenced. DNA extraction was performed as
498 previously described [16]. Library preparation and paired end sequencing using an Illumina
499 HiSeq2000 platform were conducted at Aros, Skejby, Denmark. Sequence data of the ten
500 isolates has been deposited under the NCBI BioProject IDs PRJNA312067. We used
501 SOAPdenovo2 [30] to construct *de novo* genome assemblies for each isolate independently.
502 For each genome, the k-mer value maximizing the weighted median (N50) of contigs and
503 scaffolds was selected.
504 Prior to generating a multiple genome alignment, we pre-processed the individual genomes
505 of the thirteen *Z. tritici* isolates. First, we masked repetitive sequences using a library of 497

506 repeat families identified *de novo* in four *Zymoseptoria* species [19]. Repeats were soft-
507 masked using the program RepeatMasker (option -xsmall) to retained information of repeat
508 sites in the alignment (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at
509 http://repeatmasker.org). Second, we filtered the genome assemblies to contain only contigs
510 with a length >= 1 kb. Multiple genome alignments were generated by the MULTIZ program
511 using the LASTZ pairwise aligner from the Threaded Blockset Aligner (TBA) package [69]. The
512 alignment was projected on the IPO323 reference genome using the *maf_project* program
513 from the TBA package.
514
515 **Inference of population structure**
516 In order to infer population structure, we generated genealogies of the thirteen isolates using
517 the multiple genome alignment. We used the MafFilter program [70] to compute pairwise
518 distance matrices using maximum likelihood under a Kimura 2 parameter model in 10 kb
519 sliding windows along the chromosomes of the reference genome. For each window, a BioNJ
520 tree was reconstructed from the distance matrices. The resulting 1,850 genealogies were
521 used to build a super tree using SDM for the generation of a distance supermatrix [71] and
522 FastME was used to infer a consensus tree [72]. We used the program ADMIXTURE [28] to
523 estimate the number of ancestral populations based on single nucleotide polymorphism data.
524 Input files for ADMIXTURE were generated from the filtered alignment using MafFilter [70].
525
526 **Presence-absence patterns of chromosomes and genes**
527 Based on the multiple genome alignment, the numbers of aligned bases for each genome
528 were computed in 10 kb non-overlapping sliding windows along the alignment projected on
529 the reference sequence. To assess the presence-absence of protein coding genes we used a
530 recently updated annotation of the reference genome of *Z. tritici* IPO323 [20]. Nucleotide
531 sequences corresponding to genes were used as queries for BLASTn searches against each of
532 the re-sequenced genomes. The presence of each gene in the individual genomes was
533 assessed if the best BLAST hit showed at least 90% (and up to 110%) of coverage of the total
534 length of the IPO323 gene.
535
536 **Estimation of rates of adaptation**
537 Based on the coordinates of each predicted gene model in the reference genome IPO323
538 [19,20], exons were extracted from the multiple genome alignment of *Z. tritici* isolates using
539 MafFilter [70]. Complete coding sequences (CDS) were concatenated to generate individual

540 alignments of all orthologous CDS. If one or more exons were not extracted from the
541 alignment due to missing information, the gene was discarded from further analyses. Each
542 complete CDS alignment was filtered according to the following criteria: (i) CDS were
543 discarded if they contained more than 5% gaps in one or more individuals, (ii) CDS with
544 premature stop codon were likewise deleted, and (iii) only alignments comprising three or
545 more CDS were kept. In some cases, due to indels in the genome alignment, the codon
546 phasing of some genes was lost. This issue was overcome by refining the CDS alignment using
547 the codon-based multiple alignment program MACSE [73]. The final data set contains 7,040
548 gene alignments. We further created a data set containing an outgroup sequence, taken from
549 *Z. ardabiliae*, leading to 6,767 alignments. The data set with outgroup was used to infer the
550 synonymous and non-synonymous divergence based on the rate of synonymous and non-
551 synonymous substitutions. The synonymous and non-synonymous unfolded site frequency
552 spectra (SFS) were also computed, using the outgroup sequence to orient mutations. All
553 calculations were performed using the BppPopStats program from the Bio++ Program Suite
554 [74]. We used the Dfem program in order to estimate the distribution of fitness effects from
555 the SFS and compute a genome wide estimate of α and ω_a, the proportion of mutations fixed
556 by selection and the rate of adaptive substitutions respectively [12]. The following models
557 were fitted and compared using Akaike's information criterion: Neutral, Gamma Gamma-
558 Exponential, Displaced Gamma, Scaled Beta and Bessel K. Analyses were conducted on the
559 complete set of gene alignments, as well as on a subset of genes encoding effector candidates
560 only. We further stratified our data set according to the flanking recombination rate of each
561 gene in 500 bp 5' and 3' flanking regions [25]. We discretized the observed distribution of the
562 population recombination rate ρ in 15 categories with similar number of genes, each category
563 comprising between 676 and 749 genes. For each of the 15 gene sets, 10 bootstrap replicates
564 were generated by sampling genes randomly in each category. Genes in each replicate were
565 concatenated and the Dfem program run with the GammaExpo distribution of fitness effect
566 [12]. For each recombination category, the mean estimates of α and ω_a, as well as the
567 standard error over the 10 replicates, were computed and plotted (Figure 2). Kendall
568 correlation tests were performed on the 15 times 10 estimates in order to correlate both α
569 and ω_a with ρ.
570
571
572 **Model of codon sequence evolution and candidate genes under positive diversifying**
573 **selection**

574 We used all filtered CDS alignments to reconstruct genealogies for the individual genes using
575 PhyML (model HKY85) [75]. To investigate patterns of selection and infer the role of positive
576 selection on adaptive gene evolution, the program CodeML from the PAML package was used
577 [46] with the filtered multiple CDS alignments and the corresponding phylogenetic trees as
578 inputs. CodeML allows inference of selection and evolutionary rates by calculating the
579 parameter ω, the ratio of non-synonymous to synonymous rates (dN/dS) for each gene. More
580 specifically, we compared site models that allow ω to vary among codons in the protein [18].
581 The models used in this study include the nearly neutral (M1a), positive selection (M2a), beta
582 (M7), and beta&ω (M8) models. A likelihood ratio test (LRT) was used to compare the fit of
583 null models and alternative models, and the significance of the LRT statistic was determined
584 using a χ^2 distribution. The first LRT tests for the occurrence of sites under positive selection
585 by comparing the M1a and M2a models. In the model M1a sites can be under purifying
586 selection ($0 < \omega < 1$) and evolve by neutral evolution ($\omega = 0$) while the M2a model allows for
587 some sites to be under positive selection ($\omega > 1$). The second LRT compares the M7 and M8
588 models, where in M7 a discretized beta distribution for ω (limited to the interval [0,1]) over
589 all sites is assumed, while the model M8 allows sites with $\omega > 1$. By allowing for a wider range
590 of strength of purifying selection, the M7 and M8 models are more biologically realistic. The
591 M7-M8 LRT, however, was shown to display an increased false discovery rate compared to
592 the M1a-M2a comparison [47]. For these two tests, p-values were obtained with a χ^2 test (df
593 = 2). We corrected for multiple testing and a false discovery rate of 1% was used for the
594 detection of genes under positive selection [76]. Only genes significant for both tests were
595 considered as genes evolving under positive selection.
596 To address divergent adaptation, we compared gene evolution among four closely related
597 *Zymoseptoria* species. In a previous study we defined the core proteome of *Z. tritici*, *Z.*
598 *ardabiliae*, *Z. brevis* and *Z. pseudotritici* comprising 7,786 orthologous genes [20]. We
599 generated alignments of the corresponding coding sequences using the MACSE sequence
600 aligner [73] and used CodeML with a branch model that allows ω to vary among branches of
601 the phylogeny [18]. As input we applied a non-rooted tree of the four *Zymoseptoria* species as
602 published in [77]. Branch lengths were re-estimated for each gene by CodeML [46].
603
604 **Functional enrichment analysis and gene cluster analysis**
605 PFAM domains were extracted from Interproscan results from [20]. Only domain hits with e-
606 values lower than 1e-5 were considered resulting in 10,026 domains present in 7,343 genes.
607 Enrichment tests were performed based on contingency tables, counting the number of

608 genes containing the domain and the number of genes which do not contain it, for both the
609 complete proteome and a given set of candidates to test. A χ^2 test was performed to assess
610 significance.
611 To analyse the distribution of genes under positive selection, we considered two genes
612 separated by less than 5,000 bp to be clustered and assessed the probability of such clusters
613 under a random distribution of genes along the chromosomes. To do so on a genome-wide
614 scale, we calculated the probability to obtain clusters encompassing from two to ten genes
615 under positive selection when these genes are randomly distributed across all gene
616 coordinates. Based on 10,000 random permutations, it appeared that only clusters containing
617 more than three genes were significant at the 5% level.
618
619 **Genome-wide analysis of selection patterns**
620 We compared the selection regime of each gene to several variables. Local recombination
621 rate was computed using the fine-scale recombination map of Stukenbrock and Dutheil [25]
622 as the average in the 500 bp 5' and 3' flanking regions. Local gene density was computed as
623 the proportion of coding sites in 10 kb upstream and downstream regions. Similarly, the
624 density of TEs was computed as the proportion of sites annotated as TE in a windows starting
625 10 kb upstream the gene and ending 10 kb downstream the gene, allowing for TE located
626 within the gene itself. GC content at third codon position (GC3) was also computed.
627 Expression levels were calculated as in [50]. Mean expression level was computed as the
628 maximum value observed for the gene in axenic culture or plant infection, each averaged
629 over three biological replicates. Genes encoding secreted proteins were predicted by the
630 program SignalP [78]. Genes with no homolog in any other species were considered as
631 "orphan" genes and genes located on accessory chromosomes were considered as
632 "dispensable".
633 A linear model was defined with the strength of purifying selection ($\omega 0$) as a response
634 variable, together with all individual explanatory variables. We further included an interaction
635 of gene density and recombination rate, as expected under a background selection model and
636 an interaction between protein length given that orphan genes are reported to be
637 significantly shorter than other genes [48,49]. Genes with $\omega 0 < 0.001$ or $\omega 0 > 999.0$ were
638 discarded for further analyses as these estimates most likely indicate model fitting issues,
639 possibly due to lack of data. Recombination rate and protein length variables were log
640 transformed as they showed extreme distribution. Expression values were log(x+1)
641 transformed. A BoxCox transform was used to normalize the model residues. After

642 normalization, residues displayed a good fit to a normal distribution (Shapiro-Wilk normality
643 test, p-value = 0.07534), no significant heteroscedasticity (Harrison-McCabe test, p-value =
644 0.0837), and model residues were found to be independent (Box-Ljung test, p-value =
645 0.1982). A A binary logistic regression was fitted in order to predict the occurrence of genes
646 under positive diversifying selection. The result of the LRT on codon models was used as a
647 response variable, and gene expression, recombination rate, protein length, gene density,
648 prediction of secretion, location on dispensable chromosome and whether the gene is an
649 orphan or not were used as explanatory variable, and transformed where needed. We further
650 included the effect of density of transposable elements (TE). The TE density variable was
651 highly skewed with an excess of low values, it was therefore transformed into a binary
652 variable with TE density > 0.2 (1,881 genes) and TE density <= 0.2 (9958 genes). For all linear
653 models, we used the rms package in R to compute robust regression estimates and the
654 *fastbw* to select the factors that contribute significantly to the variance explained by the
655 model [79]. We selected the 10% *Z. tritici* most expressed genes and computed the relative
656 synonymous codon usage of every codons [80]. Analyses were conducted using the 'uco'
657 function of the seqinr package for R [81].
658
659 **Simulation with recombination**
660 We used the MS program in order to simulated ancestral recombination graphs (ARGs) for
661 thirteen individuals and a given recombination rate [82]. The resulting ARGs were used to
662 simulate codon alignments using the bppSeqGen program [83]. Codon sites were simulated
663 with dN/dS values of 0.1, 0.3, 0.6, 1.0 or 3.0, with equal probability. In this set-up,
664 recombination was effectively allowed only between codons and not within codons. The
665 resulting alignments were given as input to the PhyML program [84] with a HKY85
666 substitution model and a discretized gamma distribution of site-specific rates with 4 classes,
667 as used for inference of gene data sets in this study. M1a and M2a models were then fitted
668 using the estimated tree as input with CodeML. Recombination rate of ρ = 0.0, 0.001, 0.01,
669 0.05 and 0.1, for gene lengths of 100, 500, 1,000 and 2,000 codons with 30 independent
670 replicates for each combination of parameters. CodeML output was parsed using BioPython
671 [85].
672
673 **Substitution mapping**
674 In order to assess the putative occurrence of RIP mutations among mutations observed in a
675 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 [83]. 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.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material: The datasets generated and analysed during the current study are available at NCBI under the BioProject ID PRJNA312067.

Competing interests: The authors declare that they have no competing interests

Funding: The study was funded by a Max Planck fellowship and a personal grant from the State of Schleswig-Holstein, Germany both to EHS.

Authors' contributions: JD and EHS conceived and planned the experiments. JG and JD performed the computational framework and analysed the data. All authors contributed to the interpretation of data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements: The authors thank Nicolas Galtier and Thomas Bataillon for helpful discussion.

References

1. Möller M, Stukenbrock EH. Evolution and genome architecture in fungal plant pathogens. Nat. Rev. Microbiol. 2017; 15, 771.

2. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, et al. Fungal effectors and plant susceptibility. Annu. Rev. Plant Biol. Annual Reviews; 2015;66:513–45.

3. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, et al. Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism. Science (80-.). 2010;330:1543–6.

709 4. Klosterman SJ, Subbarao K V, Kang S, Veronese P, Gold SE, Thomma BPHJ, et al.
710 Comparative Genomics Yields Insights into Niche Adaptation of Plant Vascular Wilt Pathogens.
711 PLoS Pathog; 2011;7:e1002137.
712 5. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-JM-J, Di Pietro A, et al.
713 Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature;
714 2010;464:367–73.
715 6. Daverdin G, Rouxel T, Gout L, Aubertot J-N, Fudal I, Meyer M, et al. Genome Structure and
716 Reproductive Behaviour Influence the Evolutionary Potential of a Fungal Phytopathogen.
717 PLOS Pathog.; 2012;8:e1003020.
718 7. Dutheil JY, Mannhaupt G, Schweizer G, Sieber CMK, Münsterkötter M, Güldener U, et al. A
719 tale of genome compartmentalization: the evolution of virulence clusters in smut fungi.
720 Genome Biol. Evol. 2016;8:681–704.
721 8. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of
722 molecular evolution. Nature; 1977;267:275.
723 9. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and
724 nonsynonymous nucleotide substitutions. Mol. Biol. Evol. SMBE; 1986;3:418–26.
725 10. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila.
726 Nature; 1991;351:652.
727 11. Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol. Evol. Elsevier;
728 2006;21:569–75.
729 12. Galtier N. Adaptive protein evolution in animals and the effective population size
730 hypothesis. PLoS Genet.; 2016;12:e1005774.
731 13. Smith NGC, Eyre-Walker A. Adaptive protein evolution in Drosophila. Nature;
732 2002;415:1022–4.
733 14. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev
734 Genet. 2007;8:610–8.
735 15. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, et al. Genome
736 Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. Science (80-.).
737 2010;330:1540–3.
738 16. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, et al. The making of a new
739 pathogen: Insights from comparative population genomics of the domesticated wheat
740 pathogen Mycosphaerella graminicola and its wild sister species. Genome Res. 2011;21:2157–
741 66.
742 17. Charlesworth D. Balancing Selection and Its Effects on Sequences in Nearby Genome

743 Regions. PLOS Genet. 2006;2:e64.
744 18. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and
745 applications to the HIV-1 envelope gene. Genetics. Genetics Soc America; 1998;148:929–36.
746 19. Goodwin SB, M'barek S Ben, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished
747 genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome
748 structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 2011;7:e1002070.
749 20. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq Based Gene Annotation and
750 Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify
751 Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3
752 Genes|Genomes|Genetics. 2015;
753 21. Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, et al. Histone
754 modifications rather than the novel regional centromeres of Zymoseptoria tritici distinguish
755 core and accessory chromosomes. Epigenetics Chromatin. BioMed Central; 2015;8:1.
756 22. Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH. Whole-
757 Genome and Chromosome Evolution Associated with Host Adaptation and Speciation of the
758 Wheat Pathogen Mycosphaerella graminicola. PLoS Genet; 2010;6:e1001189.
759 23. Poppe S, Dorsheimer L, Happel P, Stukenbrock EH. Rapidly Evolving Genes Are Key Players
760 in Host Specialization and Virulence of the Fungal Wheat Pathogen Zymoseptoria tritici
761 (Mycosphaerella graminicola). PLoS Pathog; 2015;11:e1005055.
762 24. Croll D, Lendenmann MH, Stewart E, McDonald BA. The Impact of Recombination
763 Hotspots on Genome Evolution of a Fungal Plant Pathogen. Genetics 2015;201:1213–28.
764 25. Stukenbrock EH, Dutheil JY. Comparison of fine-scale recombination maps in fungal plant
765 pathogens reveals dynamic recombination landscapes and intragenic hotspots. bioRxiv. 2017;
766 26. Linde CC, Zhan J, McDonald BA. Population Structure of Mycosphaerella graminicola:
767 From Lesions to Continents. Phytopathology; 2002;92:946–55.
768 27. Banke S, McDonald BA. Migration patterns among global populations of the pathogenic
769 fungus Mycosphaerella graminicola. Mol. Ecol. 2005;14:1881–96.
770 28. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated
771 individuals. Genome Res. Cold Spring Harbor Lab; 2009;19:1655–64.
772 29. O'Donnell K, Kistler HC, Tacke BK, Casper HH. Gene genealogies reveal global
773 phylogeographic structure and reproductive isolation among lineages of Fusarium
774 graminearum, the fungus causing wheat scab. Proc. Natl. Acad. Sci. 2000;97:7905–10.
775 30. Ali S, Gladieux P, Leconte M, Gautier A, Justesen AF, Hovmøller MS, et al. Origin,
776 Migration Routes and Worldwide Population Genetic Structure of the Wheat Yellow Rust

777 Pathogen Puccinia striiformis f.sp. tritici. PLoS Pathog; 2014;10:e1003903.
778 31. Brown JKM, Hovmøller MS. Aerial dispersal of pathogens on the global and continental
779 scales and its impact on plant disease. Science. 2002;297:537–41.
780 32. Mehrabi R, Taga M, Kema GHJ. Electrophoretic and cytological karyotyping of the foliar
781 wheat pathogen Mycosphaerella graminicola reveals many chromosomes with a large size
782 range. Mycologia. 2007;99:868–76.
783 33. Brunner PC, Torriani SFF, Croll D, Stukenbrock EH, McDonald BA. Coevolution and life
784 cycle specialization of plant cell wall degrading enzymes in a hemibiotrophic pathogen. Mol.
785 Biol. Evol. 2013;30:1337–47.
786 34. McDonald MC, McGinness L, Hane JK, Williams AH, Milgate A, Solomon PS. Utilizing Gene
787 Tree Variation to Identify Candidate Effector Genes in Zymoseptoria tritici. G3
788 Genes|Genomes|Genetics. 2016;6:779–91.
789 35. Croll D, Zala M, McDonald BA. Breakage-fusion-bridge cycles and large insertions
790 contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. PLoS
791 Genet.; 2013;9:e1003567.
792 36. Plissonneau C, Stürchler A, Croll D. The Evolution of Orphan Regions in Genomes of a
793 Fungal Pathogen of Wheat. MBio. Am Soc Microbiol; 2016;7:e01231-16.
794 37. de Jonge R, Peter van Esse H, Maruthachalam K, Bolton MD, Santhanam P, Saber MK, et
795 al. Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered
796 by genome and RNA sequencing. Proc. Natl. Acad. Sci. U. S. A.; 2012;109:5110–5.
797 38. Xue M, Yang J, Li Z, Hu S, Yao N, Dean RA, et al. Comparative Analysis of the Genomes of
798 Two Field Isolates of the Rice Blast Fungus Magnaporthe oryzae. PLOS Genet.;
799 2012;8:e1002869.
800 39. Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov M V, et al.
801 Genome wide analyses reveal little evidence for adaptive evolution in many plant species.
802 Mol. Biol. Evol.; 2010;27:1822–32.
803 40. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet. Res.;
804 1966;8:269–94.
805 41. Marais G, Charlesworth B. Genome Evolution: Recombination Speeds Up Adaptive
806 Evolution. Curr. Biol.; 2003;13:R68–70.
807 42. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends
808 Ecol. Evol. 2014;29:51–63.
809 43. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background
810 selection. Genet. Res. 1996;67:159–74.

811 44. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, et al. Comparative
812 population genomics in animals uncovers the determinants of genetic diversity. Nature.
813 2014;515:261–3.
814 45. Ellegren H, Galtier N. Determinants of genetic diversity. Nat. Rev. Genet.; 2016;
815 46. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.
816 2007;24:1586–91.
817 47. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in
818 detecting adaptive molecular evolution. Mol. Biol. Evol.; 2001;18:1585–92.
819 48. Liao B-Y, Scott NM, Zhang J. Impacts of Gene Essentiality, Expression Pattern, and Gene
820 Compactness on the Evolutionary Rate of Mammalian Proteins. Mol. Biol. Evol.
821 2006;23:2072–80.
822 49. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in
823 primate lineage-specific genes. Genome Biol. Evol. 2010;2:393–409.
824 50. Kellner R, Bhattacharyya A, Poppe S, Hsu TY, Brem RB, Stukenbrock EH. Expression
825 Profiling of the Wheat Pathogen Zymoseptoria tritici Reveals Genomic Patterns of
826 Transcription and Host-Specific Regulatory Programs. Genome Biol. Evol. 2014;6:1353–65.
827 51. Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast
828 protein evolution. Mol. Biol. Evol. 2006;23:327–37.
829 52. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on
830 neutral molecular variation. Genetics. Genetics Soc America; 1993;134:1289–303.
831 53. Campos JL, Halligan DL, Haddrill PR, Charlesworth B. The relation between recombination
832 rate and patterns of molecular evolution and variation in Drosophila melanogaster. Mol. Biol.
833 Evol. 2014;31:1010–28.
834 54. Duret L. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev.
835 2002;12:640–9.
836 55. Birdsell JA. Integrating genomics, bioinformatics, and classical genetics to study the effects
837 of recombination on genome evolution. Mol Biol Evol 2002;19.
838 56. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon
839 usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. 1999;96:4482–7.
840 57. André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M. CAZyme discovery
841 and design for sweet dreams. Curr. Opin. Chem. Biol. 2014;19:17–24.
842 58. Howlett BJ. Secondary metabolite toxins and nutrition of plant pathogenic fungi. Curr.
843 Opin. Plant Biol. 2006;9:371–5.
844 59. Van den Ackerveken GFJM, Vossen P, De Wit J. The AVR9 race-specific elicitor of

845 Cladosporium fulvum is processed by endogenous and plant proteases. Plant Physiol. Am Soc
846 Plant Biol; 1993;103:91–6.
847 60. Stergiopoulos I, van den Burg HA, Ökmen B, Beenen HG, van Liere S, Kema GHJ, et al.
848 Tomato Cf resistance proteins mediate recognition of cognate homologous effectors from
849 fungi pathogenic on dicots and monocots. Proc. Natl. Acad. Sci. 2010;107:7610–5.
850 61. Cambareri EB, Jensen BC, Schabtach E, Selker EU. Repeat-induced GC to AT mutations in
851 Neurospora. Science (80-.). American Association for the Advancement of Science;
852 1989;244:1571–5.
853 62. Idnurm A, Howlett BJ. Analysis of loss of pathogenicity mutants reveals that repeat-
854 induced point mutations can occur in the Dothideomycete Leptosphaeria maculans. Fungal
855 Genet. Biol. 2003;39:31–7.
856 63. Laurie JD, Ali S, Linning R, Mannhaupt G, Wong P, Guldener U, et al. Genome Comparison
857 of Barley and Maize Smut Fungi Reveals Targeted Loss of RNA Silencing Components and
858 Species-Specific Presence of Transposable Elements. Plant Cell. 2012;24:1733–45.
859 64. Hood ME, Katawczik M, Giraud T. Repeat-induced point mutation and the population
860 structure of transposable elements in Microbotryum violaceum. Genetics. Genetics Soc
861 America; 2005;170:1081–9.
862 65. Poppe S, Dorsheimer L, Happel P, Stukenbrock EH. Rapidly Evolving Genes Are Key Players
863 in Host Specialization and Virulence of the Fungal Wheat Pathogen Zymoseptoria tritici
864 (Mycosphaerella graminicola). PLoS Pathog. 2015;11:e1005055.
865 66. Marshall R, Kombrink A, Motteram J, Loza-Reyes E, Lucas J, Hammond-Kosack KE, et al.
866 Analysis of two in planta expressed LysM effector homologs from the fungus Mycosphaerella
867 graminicola reveals novel functional properties and varying contributions to virulence on
868 wheat. Plant Physiol. 2011;156:756–69.
869 67. de Jonge R, van Esse HP, Kombrink A, Shinya T, Desaki Y, Bours R, et al. Conserved fungal
870 LysM effector Ecp6 prevents chitin-triggered immunity in plants. Science (80-.);
871 2010;329:953–5.
872 68. Kulkarni RD, Thon MR, Pan H, Dean RA. Novel G-protein-coupled receptor-like proteins in
873 the plant pathogenic fungus Magnaporthe grisea. Genome Biol. BioMed Central; 2005;6:R24.
874 69. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, et al. Aligning Multiple
875 Genomic Sequences With the Threaded Blockset Aligner. Genome Res. 2004;14:708–15.
876 70. Dutheil JY, Gaillard S, Stukenbrock EH. MafFilter: a highly flexible and extensible multiple
877 genome alignment files processor. BMC Genomics. 2014;15:53.
878 71. Criscuolo A, Berry V, Douzery EJP, Gascuel O. SDM: a fast distance-based approach for

879 (super) tree building in phylogenomics. Syst. Biol.; 2006;55:740–55.
880 72. Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-
881 based phylogeny inference program. Mol. Biol. Evol.; 2015;32:2798–800.
882 73. Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: Multiple Alignment of Coding
883 SEquences accounting for frameshifts and stop codons. PLoS One. 2011;6:e22594.
884 74. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, et al. Bio++: efficient
885 extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol.;
886 2013;30:1745–50.
887 75. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large
888 phylogenies by maximum likelihood. Syst. Biol.; 2003;52:696–704.
889 76. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
890 approach to multiple testing. J. R. Stat. Soc. Ser. B.; 1995;289–300.
891 77. Stukenbrock EH, Quaedvlieg W, Javan-Nikhah M, Zala M, Crous PW, McDonald BA.
892 Zymoseptoria ardabiliae and Z. pseudotritici, two progenitor species of the septoria tritici leaf
893 blotch fungus Z. tritici (synonym: Mycosphaerella graminicola). Mycologia. 2012;104:1397–
894 407.
895 78. Petersen TN, Brunak S, Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides
896 from transmembrane regions. Nat Methods. 2011;8.
897 79. Harrell F. Regression modeling strategies: with applications to linear models, logistic and
898 ordinal regression, and survival analysis. Springer; 2015.
899 80. Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: cluster analysis clearly
900 differentiates highly and lowly expressed genes. Nucleic Acids Res. 1986;14:5125–43.
901 81. R Core Team. R: A language and environment for statistical computing. Vienna, Austria;
902 2013.
903 82. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation.
904 Bioinformatics. 2002;18:337–8.
905 83. Dutheil J, Boussau B. Non-homogeneous models of sequence evolution in the Bio++ suite
906 of libraries and programs. BMC Evol. Biol. 2008;8:1.
907 84. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online—a web server for fast
908 maximum likelihood-based phylogenetic inference. Nucleic Acids Res.; 2005;33:W557–9.
909 85. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely
910 available Python tools for computational molecular biology and bioinformatics.
911 Bioinformatics. 2009;25:1422–3.
912

Tables

Table 1: Estimates of the proportion of adaptive mutation (α) under various models of distribution of fitness effects with the DFEM method. CEG: candidate effector genes, α: proportion of adaptive substitutions, ω_a: rate of adaptive substitutions.

Data set	Model	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
CEG	Neutral	16	-102,64	237,28	0,32	0,07
CEG	Gamma	17	-90,05	214,1	0,52	0,11
CEG	GammaExpo	19	-84,6	**207,2**	**0,68**	**0,14**
CEG	DisplacedGamma	18	-90,41	216,82	0,53	0,11
CEG	ScaledBeta	18	-86,69	209,38	0,6	0,12
CEG	BesselK	19	-86,41	210,82	0,5	0,1

Table 2: Genomic factors affecting the strength of purifying selection.

Effect	Coefficient	S.E.	P-value
Intercept	2,4900	0,3580	<0.0001
Expression	-0,0602	0,0140	<0.0001
Recombination rate	-0,2030	0,0100	<0.0001
Protein size	-0,3749	0,0353	<0.0001
GC 3rd codon position	-5,2429	0,3692	<0.0001
Orphan	0,6175	0,1401	<0.0001

Table 3: Genomic factors and occurrence of positive diversifying selection.

Effect	Coefficient	S.E.	P-value
Intercept	-8,8321	0,6100	<0.0001
Expression	-0,3574	0,0388	<0.0001
Recombination rate	0,1991	0,0284	<0.0001
Protein size	1,3408	0,0954	<0.0001

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.

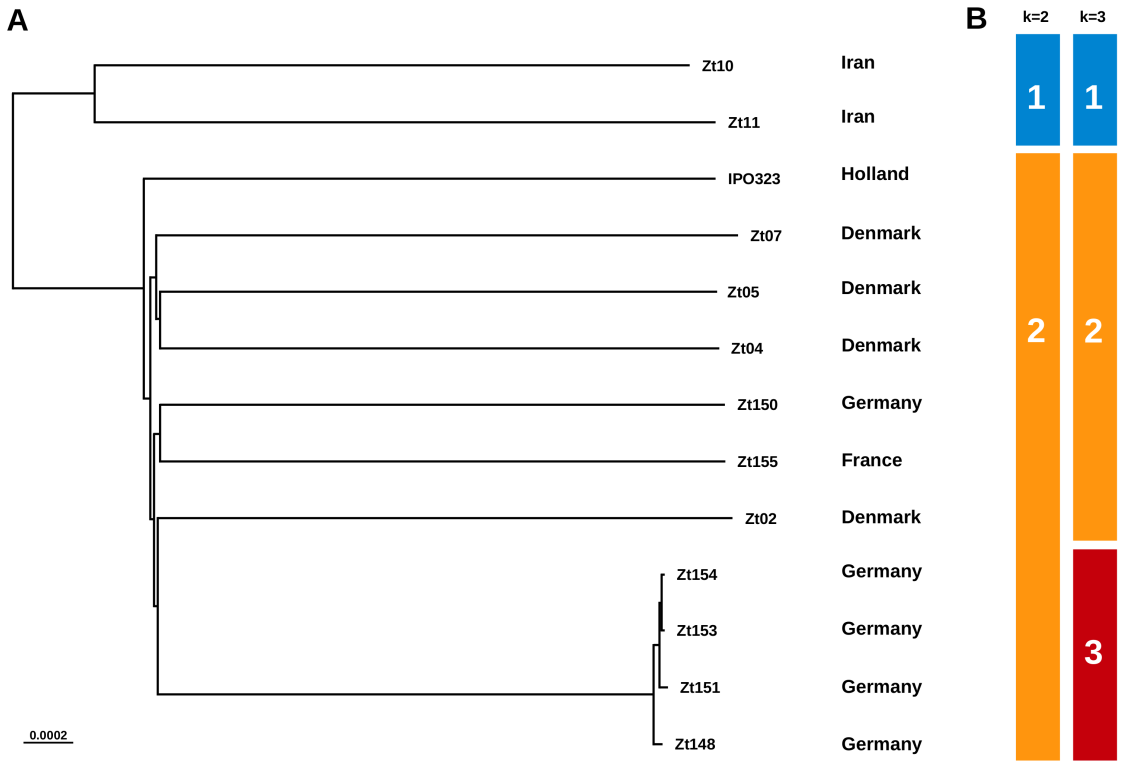
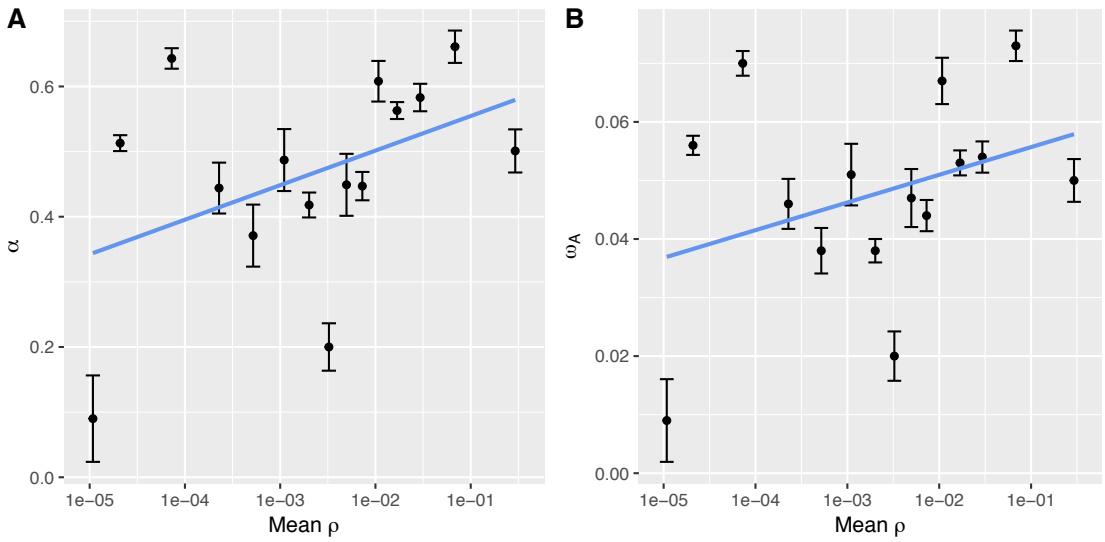
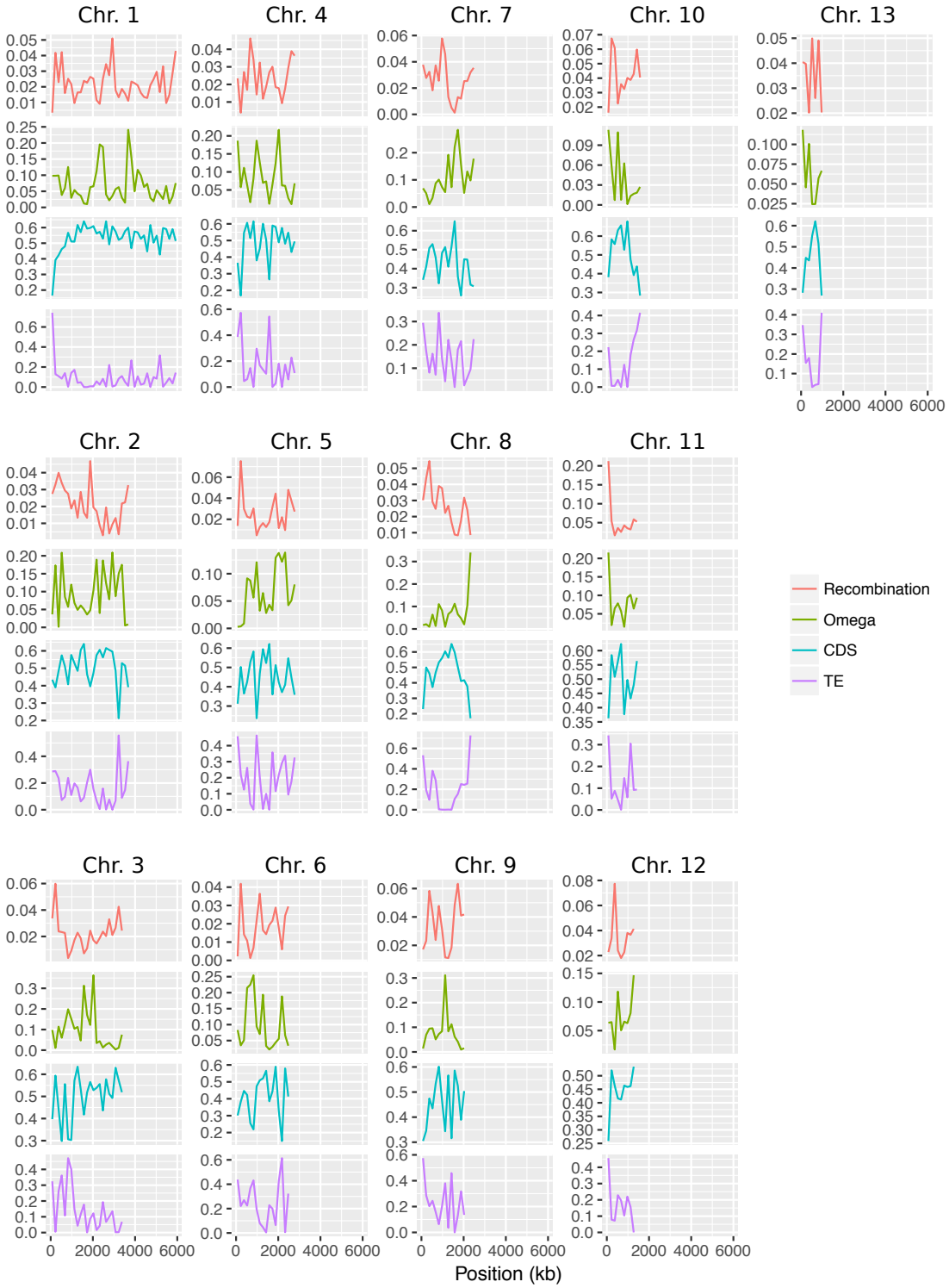


Figure 2: Estimates of A) the proportion of adaptive substitution α, and B) the rate of adaptive substitution, ω_a as a function of the population recombination rate (ρ). Each point and bars represent the mean estimate and corresponding standard error for one recombination category over 10 bootstrap replicates. The blue lines show the fitted linear models.

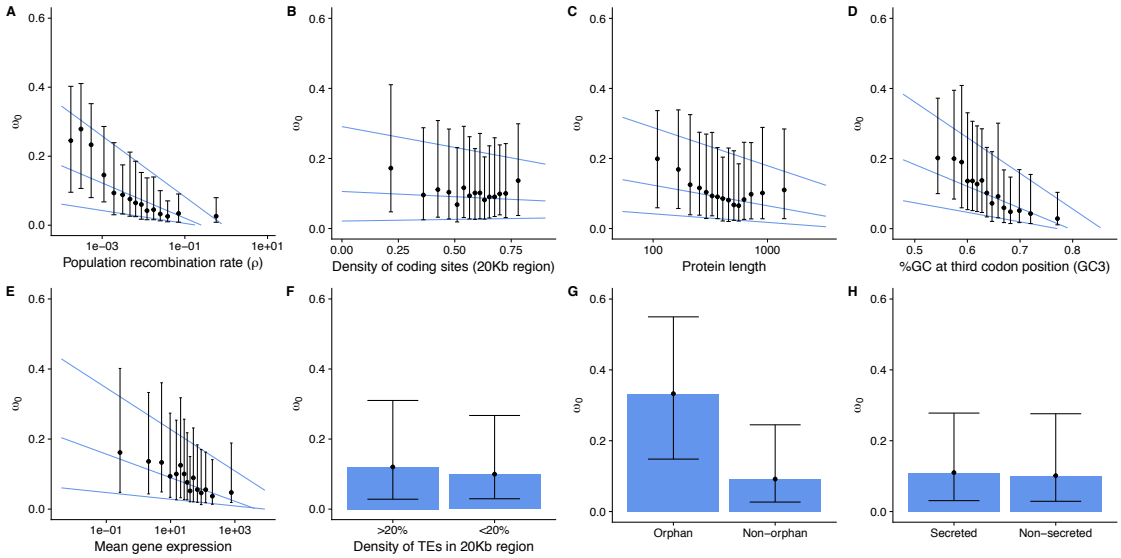


948 **Figure 3: Patterns of selection along the genome of Z. *tritici*.** Recombination, intensity of
949 purifying selection (as estimated from the ω0 parameter, see Methods), density of coding
950 sites (CDS) and density of transposable elements (TE) are plotted in windows of 150 kb along
951 the thirteen essential chromosomes.

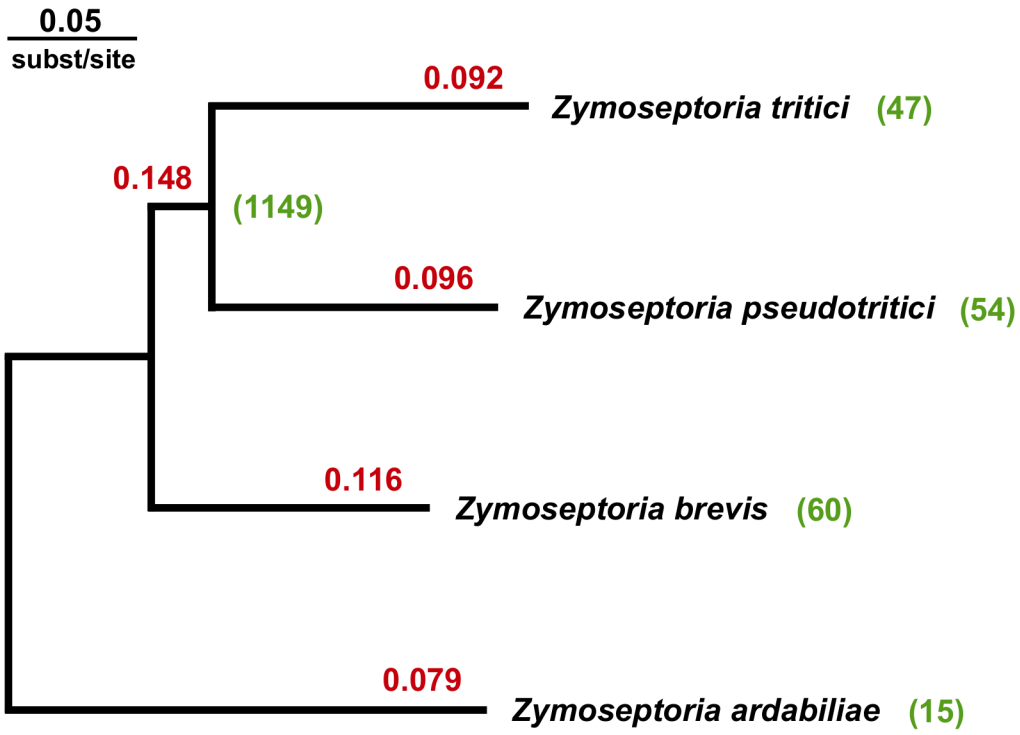


952
953

Figure 4: Correlation of the strength of purifying selection with several genomic factors. The intensity of purifying selection is measured by the ω0 parameter (see Methods). TE = transposable elements. Points represent median values and error bar the first and third quartiles of the distributions. A-E: x-axis were discretized in categories with equal point densities for clarity of visualization. Lines represent first, median and third quantile regression on non-discretized data.



985 **Figure 5: Branch-specific signatures of divergent adaptation in *Zymoseptoria* species.**
986 Phylogeny of four *Zymoseptoria* species based on integrating the mean dN / dS ratio (in red)
987 and the number of positively selected genes (in green) estimated from the 7,786 genes
988 shared by the species using the branch model implemented in PAML [46].
989
990
991
992



993
994
995
996

997 # Supplementary materials

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

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

1000 **Table S3:** Output of the PAML analysis using codon site models for the 9,412 filtered CDS of *Z.*
1001 *tritici*.

1002 **Table S4:** Functional enrichment analysis using PFAM domains for the 786 genes with sites
1003 under positive selection in *Z. tritici*.

1004 **Table S5:** Representation of RIP mutations in genes according to their distance to
1005 transposable elements.

1006 **Table S6:** Functional enrichment analysis using PFAM domains for the genes under positive
1007 selection in four *Zymoseptoria* species.

1008

1009 **Figure S1: Chromosome polymorphism in the twelve re-sequenced genomes compared to**
1010 **the IPO323 reference genome.** Based on the multiple genome alignment, for each re-
1011 sequenced genomes the alignment coverage was calculated in 10 kb sliding windows along
1012 the total length of each chromosome using the genome of IPO323 as a reference.
1013 Chromosomes 1 to 13 (in blue) are well conserved among the different isolates and constitute
1014 the "core" chromosomes while the chromosomes 14 to 21 (in red) appear to be absent or
1015 incomplete at different frequencies.

1016 **Figure S2: Impact of recombination on the estimation of the strength of purifying selection**
1017 **using phylogenetic models of codon sequence evolution.** Parameters are the ratio of
1018 transitions over transversions (kappa), the ratio of the rates of synonymous over non-
1019 synonymous mutations for sites evolving under purifying selection (ω0) in the two models M1
1020 and M2, and the ratio of the rates of synonymous over non-synonymous mutations for sites
1021 evolving under positive selection (ω1) for model M2 (see methods). Rho: scaled
1022 recombination rate used in the simulations. Length: length of the gene used in the simulation.
1023 X-axis: average number of distinct topologies per simulated site, increasing with the
1024 recombination rate. Y-axis: parameter estimate.

1025

1026 **Figure S3: Codon usage in *Z. tritici*.** Relative synonymous codon usage (RSCU) in the 10% most
1027 expressed genes of *Z. tritici*. Codon usage, according to the base type at the third position.