Genome compartmentalization predates species divergence in the plant pathogen genus *Zymoseptoria*

Alice Feurtey[1,2]*, Cécile Lorrain[1,2,3]*, Daniel Croll[4], Christoph Eschenbrenner[1,2], Michael Freitag[5], Michael Habig[1,2], Janine Haueisen[1,2], Mareike Möller[1,2,5], Klaas Schotanus[1,2,6], and Eva H. Stukenbrock[1,2]

Affiliations:

[1] Environmental Genomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany

[2] Environmental Genomics, Christian-Albrechts University of Kiel, 24118 Kiel, Germany

[3] INRA Centre Grand Est - Nancy, UMR 1136 INRA/Universite de Lorraine Interactions Arbres/Microorganismes, Champenoux 54280, France

[4] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland.

[5] Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR, USA

[6] Department of Molecular Genetics and Microbiology, Duke University, Duke University Medical Center, Durham, NC 27710, USA

*Authors contributed equally to the work. Correspondence: Alice Feurtey (feurtey@evolbio.mpg.de), Cécile Lorrain (lorrain@evolbio.mpg.de)

Abstract

Background: Antagonistic co-evolution can drive rapid adaptation in pathogens and shape genome architecture. Comparative genome analyses of several fungal pathogens revealed highly variable genomes, for many species characterized by specific repeat-rich genome compartments with exceptionally high sequence variability. Dynamic genome architecture may

enable fast adaptation to host genetics. The wheat pathogen *Zymoseptoria tritici* with its highly variable genome and has emerged as a model organism to study the genomic evolution of plant pathogens. Here, we compared genomes of *Z. tritici* isolates and genomes of sister species infecting wild grasses to address the evolution of genome composition and structure.

Results: Using long-read technology, we sequenced and assembled genomes of *Z. ardabiliae*, *Z. brevis*, *Z. pseudotritici* and *Z. passerinii*, together with two isolates of *Z. tritici*. We report a high extent of genome collinearity among *Zymoseptoria* species and high conservation of genomic, transcriptomic and epigenomic signatures of compartmentalization. We identify high gene content variability both within and between species. In addition, such variability is mainly limited to the accessory chromosomes and accessory compartments. Despite strong host specificity and non-overlapping host-range between species, effectors are mainly shared among *Zymoseptoria* species, yet exhibiting a high level of presence-absence polymorphism within *Z. tritici*. Using *in planta* transcriptomic data from *Z. tritici*, we suggest different roles for the shared orthologs and for the accessory genes during infection of their hosts.

Conclusion: Despite previous reports of high genomic plasticity in *Z. tritici*, we describe here a high level of conservation in genomic, epigenomic and transcriptomic signatures of genome architecture and compartmentalization across the genus *Zymoseptoria*. The compartmentalized genome may reflect purifying selection to retain a functional core genome and relaxed selection on the accessory genome allowing a higher extent of polymorphism.

Keywords: *genome evolution, orphan genes, effectors, genome architecture, accessory genes*

Introduction

Co-evolution between plants and pathogens can drive rapid evolution of genes involved in antagonistic interaction [1]. In filamentous plant pathogens, rapid evolution may be fueled by highly dynamic genome architecture involving repeat-rich compartments such as gene-sparse islands of repetitive DNA and accessory chromosomes [2, 3]. These compartments can show a high plasticity revealed by a high extent of gene and/or chromosome presence-absence

variation and structural variants, such as inversions, insertions and deletions [4, 5]. Several plant pathogenic fungi have isolate specific chromosomes, so-called accessory chromosomes.

Accessory chromosomes are characterized by intra-species presence-absence polymorphism, low gene density, an enrichment of repetitive sequences and, in some species, a different histone methylation pattern [6, 7]. It has been shown that accessory chromosomes encode genes involved in virulence such as in the species *Fusarium solani*, *Fusarium oxysporum* and *Leptosphaeria maculans* [8–11]. Little is known about the evolutionary origin of accessory chromosomes although experimental evidence from the asexual species *F. oxysporum* shows that accessory chromosomes may be acquired horizontally as chromosomes can be transferred between distinct isolates by hyphal fusion [10]. Through such transfers, virulence determinants may be exchanged between clonal lineages as accessory chromosomes in this species were shown to encode host specific virulence determinants and transcription factors regulating their expression [12].

Genes involved in plant-pathogen interactions may diversify at a higher rate in repeat-rich genome compartments and thereby evolve new virulence specificity faster [3]. These genes encode secreted proteins, so-called effectors [1]. Most known effectors target diverse cellular compartments and molecular pathways, including immune response-related pathways [13, 14]. Genes encoding Carbohydrate-active enzymes (CAZymes) have also been associated to the pathogenic lifestyle of fungal plant pathogens, particularly through their role in plant-cell wall degradation [15]. Thus, some secreted CAZymes may be essential from the early infection stage, like penetration of plant tissue, to later stages such as the necrotrophic phase where the pathogen feed from dead plant tissue [16]. Likewise, secondary metabolites are known to be involved in plant infection and contribute to virulence and the interaction with other plant-associated microorganisms [17, 18]. Many of these genes can be predicted either according to their composition and known protein domains or through machine learning methods [19]. Thereby, in-depth genome annotations have proven important to predict and compare the

content of pathogenicity-related genes in plant pathogens, as well as their genomic localization for example in rapidly evolving genome compartments.

The ascomycete pathogen *Zymoseptoria tritici* has emerged as a model organism in evolutionary genomics of pathogens. This species originated in the Fertile Crescent during the domestication of its host, wheat [20]. Closely related species of *Z. tritici* have been collected from wild grasses in the Middle East providing an excellent resource for comparative genome analyses of closely related and recently diverged pathogen species. Comparative analyses of genome organization and gene content within and among *Zymoseptoria* species have previously revealed a wide distribution of accessory chromosomes and dynamic gene content [21, 22]. The haploid genome of the reference isolate IPO323 comprises thirteen core and eight accessory chromosomes [23]. Some of these accessory chromosomes encode traits that impact virulence of the fungus, however no gene encoded on an accessory chromosome has so far been described as a virulence or avirulence determinant [24–29]. Interestingly, the accessory chromosomes in *Z. tritici* show a low transcriptional activity *in vitro* as well as *in planta* [30, 31]. This suppression of gene expression correlates with an enrichment of heterochromatin associated with the histone modification H3K27me3 on the accessory chromosomes [6, 32].

In the reference isolate IPO323, the accessory chromosomes comprise more than 12% of the entire genome assembly. To which extent such a high amount of accessory DNA is also found in genomes of other members of the *Zymoseptoria* genus has so far been unknown due to the lack of high-quality genome assemblies and large scale population sequencing. Assemblies based on short-read data failed to recover complete sequence of accessory chromosomes and “orphan regions” due to their high repeat content [22]. The asset of genome assemblies based on long-read sequencing was demonstrated in a detailed genome comparison of four *Z. tritici* isolates sequenced with PacBio long-read sequencing [28]. Comparison of the high-quality chromosome assemblies revealed the occurrence of “orphan regions” enriched with transposable elements and encoding putative virulence-related genes [33].

In this study, we investigate the genomic architecture and variability among five *Zymoseptoria* genomes. Beside presenting a new and significantly improved resource for future genomic studies of these fungal pathogens, we specifically ask: 1) how conserved is the genome architecture among *Zymoseptoria* species, 2) can we identify accessory compartments in other *Zymoseptoria* isolates and 3) to which extent does variation in genome architecture reflect variation in gene content.

To answer these questions, we used high-quality assemblies based on long-read sequence data and new gene predictions in two isolates of *Z. tritici* (Zt05 and Zt10) and one isolate of each of the sister species, *Z. ardabiliae, Z. brevis, Z. passerinii,* and *Z. pseudotritici.* We explore the core and non-core genome architecture of *Zymoseptoria* spp. combining genomic data with transcriptome and histone methylation data and relate this to core and accessory genome compartments. Furthermore, we compare the distribution of orthologous and non-orthologous genes in the *Zymoseptoria* genomes and one additional Dothideomycete species. Our analyses reveal an overall conserved genome architecture characterized by gene-rich core compartments and accessory compartments enriched in species-specific genes. Finally, we report an exceptionally high extent of variation in presence-absence of protein coding genes in a eukaryote genome.

Results

New de novo assemblies using long-read sequencing for six *Zymoseptoria* spp.

We sequenced and assembled the genome of the reference isolates of *Z. ardabiliae*, *Z. brevis*, *Z. pseudotritici* and *Z. passerinii* and the genomes of two *Z. tritici* isolates sampled in Denmark and Iran. The obtained contigs were filtered based on base-quality confidence and read depth to ensure high quality of the final assemblies (see Methods). This filter removed a high number of contigs (between 17% and 58% of the total), but little overall length (between 0.4 and 2.6% of the total assemblies), indicating that most of the excluded contigs were of small size. The best

137 assemblies were of the two *Z. tritici* isolates comprising 19 and 30 contigs and the most
138 fragmented was of *Z. passerinii* comprising 103 contigs (Figure 1). The resulting assembly
139 lengths ranged from 38.1 Mb for *Z. ardabiliae* to 41.6 Mb for *Z. brevis*, which is comparable to
140 the reference assembly length of *Z. tritici* (39.7 Mb) but slightly larger than previous short-read
141 based assemblies (Table 1; previous assemblies ranged from 31.5 *Z. ardabiliae* to 32.7 for *Z.*
142 *pseudotritici* [22, 23]. The assembly of the Iranian *Z. tritici* isolate Zt10 has telomeric repeats at
143 the end of all contigs, indicating that each chromosome is completely assembled, comprising six
144 accessory and thirteen core chromosomes. The assemblies for the Danish *Z. tritici* isolate
145 (Zt05), *Z. brevis* (Zb87) and *Z. pseudotritici* (Zp13) contained, respectively, twelve, nine and five
146 fully assembled chromosomes including both core and accessory chromosomes (Figure S1).
147 The assemblies of the *Z. ardabiliae* (Za17) and *Z. passerinii* (Zpa63) genomes included no fully
148 assembled chromosomes, but twelve and ten contigs respectively with telomeres at one of the
149 ends (Table 1; Figure S1).
150 The transcriptome-based gene predictions for these new assemblies include between 10,528
151 and 12,386 protein-coding genes (Table S1). This range is consistent with the annotation of the
152 reference genome IPO323 reporting 11,839 protein-coding genes [22]. We used Benchmarking
153 Universal Single-Copy Orthologs (BUSCO) from the lineage dataset *Pezizomycotina* to evaluate
154 the completeness of the assemblies and gene predictions (Waterhouse et al. 2018). The
155 proportion of complete BUSCOs were as follow Zt10: 98.7%, Zt05: 98.4%, Zp13: 98.5%, Zb87:
156 97.0%, Za17: 98.2% and Zpa63: 97.5%. These values are comparable to the one obtained for
157 the reference genome of *Z. tritici* (97.8%). The assessment of gene content completeness (see
158 Methods) indicates that, despite more fragmented assemblies of *Z. ardabiliae* and *Z. passerinii*,
159 the genomes are complete in terms of gene content and that the unassembled fragments are
160 more likely to comprise repeats and not protein coding genes.
161 Based on the whole-genome sequences and the predicted genes we reconstructed the
162 phylogeny of the *Zymoseptoria* genus using the publicly available genome of *Cercospora*
163 *beticola* as an outgroup [34, 35]. For both trees, the phylogenetic relationship of the

Zymoseptoria species is in accordance with previously published phylogeny based on seven loci sequenced in multiples isolates (Figure 1) [21].

Genomes of *Zymoseptoria* spp. comprise accessory chromosomes and compartments but show overall high synteny

Next we addressed the extent of co-linearity of the *Zymoseptoria* genomes. Using coordinates of orthologous genes, we were able to reveal a high extent of synteny conservation among the five *Zymoseptoria* species and between the three isolates of *Z. tritici*, as depicted in Figures 2 and S2. Based on this high extent of synteny and the prediction of telomeric repeats, we identified the correspondence of chromosomes between the reference genome of *Z. tritici* IPO323 and the other *Zymoseptoria* genomes (Figure 2 and S2). *Z. brevis* and *Z. pseudotritici* share a near perfect synteny in their core chromosomes, however, when compared to *Z. tritici*, *Z. brevis* and *Z. pseudotritici* have two large-scale inversions comprising roughly ~900 kb and ~1.2 Mb of chromosomes 2 and 6, respectively (Figure 2, S2 and S3). Based on the phylogeny in Figure 1, it is likely that these two events occurred after the divergence of *Z. tritici* from *Z. brevis* and *Z. pseudotritici*. Overall, we observe a higher extent of synteny conservation between *Z. brevis* and *Z. pseudotritici* compared to *Z. tritici* IPO323 (Figure 2; S2 and S3).

In *Z. tritici*, core and accessory compartments have very distinct genomic features. It was previously shown that hallmarks of accessory regions in the reference isolate IPO323 include lower gene density, lower levels of H3K4 methylation levels and lower gene expression [6, 31]. In the reference genome of IPO323, compartments with these genomic and epigenomic hallmarks represent either accessory chromosomes or specific regions of the core chromosomes. Here we find that the specific accessory hallmarks including low gene density, low expression, low H3K4me2 methylation and significant enrichment of species-specific genes (see description below) on the non-core contigs are found in genomic compartments throughout the genus (Table S3, Figures 3 and S4).

In the reference *Z. tritici* strain, the specific “accessory-like” pattern includes a particular region of the core chromosome 7 of ~0.6 Mb (Figure 3A) [6]. This particularly large accessory marked

191 region is observed in several *Zymoseptoria* spp (Figure 3B; S4). We identify ~0.7 Mb of the
192 contig 28 in *Z. pseudotritici* and ~0.6 Mb of the contig 17 for *Z. brevis*, both corresponding to
193 chromosome 7 of *Z. tritici*, which share the same hallmarks of accessory chromosomes (Figures
194 3 and S4, Table S3). For the two remaining sister species, the fragmentation of the assembly
195 does not allow the identification of such pattern although there is an indication of a similar effect
196 on gene expression and species-specific gene enrichment on contig 19 of *Z. ardabiliae* which
197 corresponds to a fragment of chromosome 7 in IPO323 (Figure S5).
198 We also identified other regions enriched in isolate-specific genes, thus defining orphan loci in
199 the core chromosomes of both Zt10 and Zt05 (Figures 3 and S4). We observe a region of ~0.2
200 Mb of contig 1 in the Iranian isolate Zt10 corresponding to the core chromosome 3 in the IPO323
201 genome with high content of isolate-specific genes (Figure S4; table S3). We furthermore
202 identified small segments with species-specific genes on the core chromosomes of the wild-
203 grass infecting sister species including a ~0.3 Mb region of the contig 26 in *Z. brevis* and ~0.1
204 Mb of contig 30. Overall, we show that genome compartmentalization in core and accessory
205 regions is an ancestral and shared trait among the *Zymoseptoria* species. This phenomenon
206 generates highly variable compartments and defined loci that deviate from genome averages in
207 terms of gene content, sequence composition and synteny conservation.

208 **Variable repertoires of effector candidate genes**

209 To obtain gene annotations for the *Zymoseptoria* genome assemblies, we established a custom
210 pipeline adapted from Lorrain and co-workers (Figure S5) [36]. Briefly, we use the consensus of
211 three methods to predict gene product localization, then extract secreted proteins to further
212 identify predicted effectors. This detailed functional annotation provided a catalog of predicted
213 gene functions and cellular localizations (Figure S6). For each genome, a large proportion of
214 genes could not be assigned to a protein function and lacked protein domains. 49.6% of genes
215 in *Z. tritici* (N = 5953) and up to 71.8% of genes in *Z. pseudotritici* (N = 8373) lack a predicated
216 function (Figure S6A). A relatively consistent number of genes are predicted for each functional
217 category among *Zymoseptoria* spp. (Figure S6A and B). Likewise, the numbers of gene products

predicted to belong to the different subcellular localizations are very similar (Figure S6C) across the whole genus, including secreted proteins. The difference between the minimal and maximal gene number fpr the different categories of subcellular localizations do not exceed 1.6X between species (Figure S6C). Overall, secretomes range from 7% of the genes predicted in *Z. passerinii* (N=828) to 11% of genes in *Z. ardabiliae* (N=1328 genes, Figure S6B).

We further investigated the number and distribution of genes predicted to encode proteins with a pathogenicity-related function, such as secondary metabolites, CAZymes and effector candidates (Figures S1 and S6B). Genes involved in the synthesis of secondary metabolites are typically organized in clusters, with genes participating in the same biosynthetic pathway grouping together at a genomic locus (Shi-Kunne et al. 2019). The number of biosynthetic gene clusters (BGC) ranges from 25 in *Z. ardabiliae* and *Z. passerinii* to 33 in the IPO323 reference genome and includes from 305 to 471 predicted genes (Figure S1). The only BGC identified in a non-core contig is a non-ribosomal peptide synthetase BGC found on the contig 38 of *Z. brevis* which has no orthologous cluster detected in any of the other *Zymoseptoria* genomes (Figure S1). We identified between 454 and 515 CAZyme genes in the *Zymoseptoria* species. Both BGCs and CAZymes are almost exclusively found on the core chromosomes (Figure S1). The only exceptions are a CAZyme encoding gene found on chromosome 14 in *Z. tritici* IPO323 and Zt05, and a CAZyme encoding gene on the putative accessory contig 38 of *Z. brevis* (Figure S1). These two genes encode for a beta-glucosidase and a carboxylic-ester hydrolase, respectively.

In contrast to the high conservation of CAZyme and BGC gene content among the *Zymoseptoria* genomes, we find that predicted effector genes exhibit a large variation in gene numbers between genomes (Figure S6B). In fact, the predicted effector gene repertoire of in *Z. ardabiliae* (N=637) is three times as high compared to *Z. brevis* (N=206). Interestingly, the three *Z. tritici* isolates also vary considerably in their predicted effector repertoires. The reference isolate IPO323 has a reduced set of effector genes (N=274) compared to Zt05 and Zt10 that encode approximately 30% more effector genes (N=417 and N=403, respectively, Figure S6B). Despite the high variability, the effector genes are mostly located on core chromosomes and none of the

246 five *Zymoseptoria* species have more than ten effector genes located on accessory
247 chromosomes (Figure S1).

248 **The accessory genes of *Z. tritici* are shared with the closely related wild-grass infecting**
249 **species**

250 To further characterize variation in gene content among the five *Zymoseptoria* species, we
251 identified orthologous genes (i.e. orthogroups) from the gene predictions. We categorized 22341
252 gene orthogroups identified in the seven *Zymoseptoria* genomes and in *C. beticola* according to
253 their distribution among fungal genomes (Figure 4A). The core orthogroups, which are genes
254 present in all eight genomes, represent around 30% of all orthogroups (N = 6698). The *genus-*
255 *specific* orthogroups, shared between several *Zymoseptoria* spp. but not found in the *C. beticola*
256 genome, represent 45% of the orthogroups (N = 9955; ranging from 2066 to 3212 per species).
257 Among the *genus-specific* orthogroups, 1100 are found in all *Zymoseptoria* genomes (Figure
258 4A), whereas all others show presence-absence polymorphisms within the genus. A total of
259 2476 *species-specific* orthogroups (ranging from 552 to 1191 per species) are found only in
260 individual species. Among the *species-specific* genes, 205 orthogroups (N_{genes} = 414 to 562) are
261 found in all three *Z. tritici* genomes while the *isolate-specific* genes in *Z. tritici* represent 391
262 (Zt10) to 792 (IPO323) genes.
263 Comparing the three *Z. tritici* isolates independently from the other species, we observe
264 extensive gene presence-absence polymorphisms between the three isolates: 1540 orthogroups
265 are identified in only two strains and 2522 are found in only one (Figure 4B). The number of
266 genes showing presence-absence variation is striking compared to the 10098 core genes in *Z.*
267 *tritici* as these genes comprise almost 30% of all predicted genes. Interestingly, we show that the
268 number of orthogroups detected as *isolate-specific* is much larger when the comparison includes
269 only members of the same species than when the other species are included (1035, 849 and
270 638 vs 792, 659 and 391 genes for IPO323, Zt05 and Zt10 respectively; Figures 4A and B). This
271 indicates that a large part of the accessory gene content in *Z. tritici* is shared among the sister

272 species, and highlight the importance of including sister species when establishing core and
273 accessory gene content.
274 Interestingly, we show that effectors are enriched among the *genus-specific* genes but not
275 among the *species-specific* or *isolate-specific* gene categories (with the exception of *Z.*
276 *ardabiliae*). Fifty-six percent of effectors in *Z. ardabiliae* and up to 78% of effectors in *Z.*
277 *pseudotritici* are shared with at least one of the other five *Zymoseptoria* species (Figure 4C).
278 Indeed, 427 effector orthogroups are found in at least two genomes. However, only 47 (10% of
279 the total effector orthogroups N= 474) are found in all seven *Zymoseptoria* genomes. Among the
280 effectors shared by *Z. tritici,* and at least one other *Zymoseptoria* species, 32% (N = 112 of 352)
281 are present in all three *Z. tritici* isolates while 68% (N = 240 of 352) show presence-absence
282 polymorphisms in at least one of the three isolates. These results indicate that the majority of
283 these shared effectors is actually accessory (i.e. presence-absence polymorphism) in *Z. tritici*.

284 **Among *in planta* differentially expressed genes, species-specific are more expressed**
285 **than core genes**

286 Finally, we addressed the functional relevance of accessory and orphan genes in *Z. tritici* by
287 analyzing gene expression patterns. We used previously published transcriptome data of three
288 *Z. tritici* isolates [30] and focused on gene expression of the above-defined categories (*core*
289 *genes, genus-specific, species-specific, and isolate-specific*). We sorted *in planta* expression
290 data into two different infection phases: the biotrophic phase and the necrotrophic phase, a
291 separation supported by principal component analysis of normalized DESeq2 counts (Figure
292 S7). We compared expression levels by mapping RNA-seq reads to the genomes of IPO323,
293 Zt05 and Zt10, using normalized read mappings to transcript per million. We tested differences
294 among gene categories using pairwise comparisons with a Kruskal-Wallis test (Figure 5).
295 Overall, we find that gene expression of the *species-specific* and *isolate-specific* genes is
296 significantly lower in IPO323 and Zt10, but not in Zt05 (Kruskal-Wallis p-value < 0.05). *Species-*
297 *specific* and *isolate-specific* gene median expression ranges from 3.2 to 5.6 TPM in IPO323 and
298 Zt10 while median expression of core genes is 12.1 and 10.9, respectively. The Zt05 expression

299 profile does not follow the same trend: the *core genes* are the lowest expressed gene category
300 (8.9 median TPM), while *genus*-; *species*- and *isolate-specific* genes showed higher transcription
301 levels (12.0; 14.4 and 13.5 median TPM respectively, Kruskal-Wallis p-value < 0.05).
302 In contrast, we observe a significantly higher expression of the *species-specific* and *isolate-*
303 *specific* genes for all three isolates (Table S2; Kruskal-Wallis p-value < 0.05) when comparing
304 the expression of genes that are differentially expressed (DEGs; DESeq2 p-adjusted < 0.05)
305 between the biotrophic and necrotrophic phases. *Species-specific* and *isolate-specific* DEGs are
306 higher expressed *in planta* than the *core* and *genus-specific* genes (Figure 5B; Kruskal-Wallis p-
307 value < 0.05). The expression pattern of DEGs with different levels of specificity present a
308 consistent pattern in all three isolates (Table S2). Overall, this comparison reveals a potential
309 functional relevance of accessory genes which are up-regulated during infection of *Z. tritici*.

310 **Discussion**

311 In this study we present a new resource of high-quality whole genome assemblies and gene
312 annotations for the species *Z. ardabiliae, Z. brevis, Z. passerinii, Z. pseudotritici,* and two *Z. tritici*
313 isolates. This new dataset provides a valuable resource for detailed analyses of genome
314 architecture and evolutionary trajectories in this group of plant pathogens. Here, we conduct
315 some first comparative analyses of genome architecture and show a considerable extent of
316 variation in sequence composition during the recent evolution of *Zymoseptoria* lineages. We
317 show that genome compartmentalization and accessory chromosomes represent shared
318 ancestral traits among the pathogen species.

319

320 We identify extensive presence-absence variation of protein coding genes in genomes of five
321 *Zymoseptoria* species consistent with the variable gene repertoire already reported for *Z. tritici*
322 [28], Furthermore, the different species, share a particular genomic architecture that comprises
323 specific accessory genome compartments. In spite of this variation, we observe an overall
324 conserved synteny of the core chromosomes. In the *Zymoseptoria* genomes, we observe gene-
325 dense, actively transcribed and H3K4me2-enriched compartments covering most of the core
326 chromosomes. These compartments are clearly distinguishable from gene-sparse, non-

transcribed and H3K4me2-deprived compartments. Based on previous analyses of accessory chromosomes in *Z. tritici*, we here consider this pattern as a specific hallmark of accessory genome compartments in the genus *Zymoseptoria* and not only in *Z. tritici* [6]. We hypothesize that these compartments likely represent accessory chromosomes in the different *Zymoseptoria* species.

We also identify accessory signatures in core chromosomes, including the previously described right arm of chromosome 7 [6]. Although this region has not been reported to share the same heavy pattern of presence-absence polymorphism as the accessory chromosomes, it is worth noting that a recent study identified a large deletion in chromosome 7 likely corresponding to this region in a single *Z. tritici* isolate originating from Yemen [37]. Here we show that this homologous region also exhibits accessory compartment hallmarks in the other *Zymoseptoria* species. A fusion between a core and an accessory chromosome occurred tens of thousands of years ago and prior to species divergence. The specific genomic and epigenetic features have remained stable through speciation and evolutionary time.

In this study, we confirm previously reported genome comparisons showing that gene content in *Z. tritici* is highly variable [28]. We further extended the identification of orthologs throughout the whole *Zymoseptoria* genus. Thereby, we show that more than 25% of the genes identified as *isolate-specific* in a comparison including only *Z. tritici* isolates are actually shared with its wild-grass infecting sister species, proving that a large proportion of the accessory genome of *Z. tritici* is not specific to this species. Instead, the accessory genome content is shared among *Zymoseptoria* species. The proportion of accessory *Z. tritici* genes shared with other *Zymoseptoria* species was found to be the highest in the Iranian isolate, which is the only isolate sympatric with the four sister-species. A likely explanation for this observation would be inter-specific gene flow which would allow the different wild species to exchange genes with sympatric *Z. tritici* isolates. This new finding is consistent with recent findings from population genomic data studies revealing extensive introgression between *Zymoseptoria* species [38, 39]. It also opens

354 new perspectives for further analysis to understand how inter-specific gene flow has affected the
355 evolution of the accessory genome of *Z. tritici*.
356
357 The genes with predicted functions and, in particular, functions related to pathogenicity are
358 largely shared in *Zymoseptoria* genus. Although the lifestyles of the wild-grass infecting
359 *Zymoseptoria* are poorly understood the species most likely share major features of their
360 lifestyles. Thus, as expected, a similar CAZymes and BGC contents in all studied *Zymoseptoria*
361 species was identified. Effector prediction is often used to better understand the adaptation of
362 pathogens to their hosts [40]. However, in *Zymoseptoria*, the effectors are significantly enriched,
363 not among the isolate- or even species-specific genes, but among the genes specific to the
364 genus. If we exclude *Z. ardabiliae*, the repertoire of effectors shared at the genus level
365 represents three-quarters of all predicted effectors. This ratio suggests either that only a small
366 number of effectors confer host specificity among the different *Zymoseptoria* species or that
367 host-specificity is conferred not by presence-absence of these effectors but by regulation of their
368 expression [30]. In the *Botrytis* genus (Dothideomycetes), sister species infecting different hosts
369 share effectors with confirmed functions [41]. We hypothesize that the different specificity levels
370 reflect functional differences in the effector repertoire of *Zymoseptoria*. Effectors conserved
371 across the *Zymoseptoria* genus are core effectors, potentially targeting key plant defense
372 mechanisms common to all of their hosts [42]. In contrast, effectors which are *species-specific* or
373 *isolate-specific* might reflect host specificity and target specific pathogen recognition pathways
374 [42]. This assumption is supported by the fact that only a fraction of the *genus-specific* effectors
375 is actually shared among all *Zymoseptoria* species while the majority shows presence-absence
376 polymorphisms. *Z. ardabiliae* has been isolated from leaves of distantly related grass species in
377 Iran, including *Lolium* spp., *Elymus repens,* and *Dactylis glomerata* and potentially resulting in a
378 broader *species-specific* effector repertoire of *Z. ardabiliae* compared to the other species [21].
379
380 Consistent with a previous study [28], we found that *Z. tritici* core genes are more expressed
381 compared to accessory genes *in planta*. Core genes are more likely to hold essential functions

382 which could explain higher expression pattern during infection. Differentially expressed genes
383 that are specifically induced during the course of the infection are very likely to have functions
384 essential to pathogenicity of the fungus. Interestingly, here we show that within the differentially
385 expressed genes between the biotrophic and the necrotrophic phases of infection, *isolate*- and
386 *species-specific* genes have higher expression levels than core genes. These *isolate*- and
387 *species-specific* genes could be functionally important and regulate functions linked to infection
388 success in the biotrophic phase or to leaf colonization in the necrotrophic phase. Since these
389 genes show presence-absence polymorphisms in the genus and in the *Z. tritici* species, they
390 could represent a reservoir for possible adaptations to either host species, host cultivars or local
391 environments.
392
393 **Conclusion**
394 We investigated the genomic architecture in a genus of plant pathogens, including the
395 economically relevant wheat pathogen *Z. tritici.* Comparing genome content and genome
396 structure, we identified a large shared effector repertoire characterized by inter- and intraspecies
397 presence-absence polymorphisms. Major features of genomic, transcriptomic and epigenetic
398 compartmentalization, distinguishing accessory and core compartments, were shared among
399 wheat and wild-grass infecting *Zymoseptoria* species. We conclude that compartmentalization of
400 genomes is an ancestral trait in the *Zymoseptoria* genus.

401 **Methods and Materials**

402 **Fungal material, DNA extraction and sequencing**

403 Details regarding the individual *Zymoseptoria* isolates can be found in Table 1. For genomic data
404 we used the three *Z. tritici* isolates IPO323 (reference), Zt05 and Zt10, one *Z. ardabiliae* isolate
405 (Za17), one *Z. brevis* isolate (Zb87), one *Z. passerinii* isolate (Zpa63), and the *Z. pseudotritici*
406 isolate (Zp13).For transcriptomic and epigenomic data we used *Z. tritici* Zt09 (IPO323 ΔChr18) a
407 derivate of the reference isolate IPO323 deleted with the chromosome 18 [31].

Long read assemblies of the *Z. tritici* isolates Zt05 and Zt10 were described and published previously [30]. For DNA extraction and long read sequencing cultures of *Z. pseudotritici*, *Z. ardabiliae, Z. brevis* and *Z. passerinii* were maintained in liquid YMS medium (4 g/L yeast extract, 4 g/L malt extract, 4 g/L sucrose) at 200 rpm and 18°C. DNA extraction was conducted as previously described [26]. PacBio SMRTbell libraries were prepared using DNA extracted from single cells based on a CTAB extraction protocol [43]. The libraries were size selected with an 8-kb cutoff on a BluePippin system (Sage Science).

After selection, the average fragment length was 15 kb. Sequencing of the isolates Za17, Zb87, and Zp13 was run on a PacBio RS II instrument at the Functional Genomics Center, Zurich, Switzerland. Sequencing of the Zpa63 isolate was performed at the Max Planck-Genome-Centre, Cologne, Germany.

Genome assembly, and repeat and gene predictions

For each isolate, we assembled the genome de novo using SMRT Analysis software v.5 (Pacific Bioscience) with two sets of parameters: default parameters and “fungal” parameters. We chose the best assemblies generated by comparison of all assembly statistics produced by the software Quast such as the number of finished contigs, the size of the assembly and the N50 [44]. Summary statistics for each assembly can be found in Table 1. In order to exclude poor quality contigs from the raw assemblies, we filtered out the contigs with less than 1.5X and more than 2X median read coverage as these might be unreliable from lack of data or because they contain only repeated DNA. This filter removed a high number of contigs, i.e., between 58% and 17% of contigs. Telomeric repeats (“CCCTAA”) were identified in the remaining contigs using bowtie2 to identify the number of fully assembled chromosomes or chromosomes arms based on the presence of more than six repeats at the contig extremities [45–47].

We next used the REPET package to annotate the repeat regions of *Z. ardabiliae* Za17, *Z. brevis* Zb87, *Z. pseudotritici* Zp13, *Z. passerinii* Zpa63, and the three *Z. tritici* isolates IPO323, Zt5 and Zt10 (https://urgi.versailles.inra.fr/Tools/REPET; [48, 49]. For each genome, we

434 annotated the repetitive regions as follows: we first identified repetitive elements in each genome
435 using TEdenovo following the developer's recommendations and default parameters. The library
436 of identified consensus repeats was then used to annotate the respective genomes using
437 TEannot with default parameters.
438 We used several different approaches to predict gene coordinates and leveraged the information
439 contained in previously published RNA sequencing data to increase the quality of the prediction
440 [22, 30, 32]. We used GeneMark-ES for the first prediction *ab initio* using the option "--fungus"
441 [50]. We furthermore used previously obtained RNAseq data in two different ways: first by
442 mapping the reads to the respective genomes and second by assembling them *de novo* into
443 longer transcripts. For this, we first trimmed the reads using Trimmomatic [51], and then mapped
444 them on the newly assembled genomes using hisat2 [47]. Next we used the BRAKER1 pipeline
445 to predict genes for each genome using the fungus flag and based on the previously mapped
446 reads [52]. BRAKER applies GeneMark-ET and Augustus to create the first step of gene
447 predictions based on spliced alignments and to produce a final gene prediction based on the
448 best prediction of the first set [53, 54]. In addition to the *ab initio* gene predictions, this produced
449 a second set of predicted genes. Furthermore, the RNAseq reads were separately assembled
450 into gene transcripts using Trinity [55]. These were aligned using PASA and EVidence Modeler
451 to produce consensus gene models from the two independent predictions and the *de novo*
452 assembled transcripts [56]. Gene counts, length and other summary statistics presented in Table
453 S1 were obtained using GenomeTools [57] and customs scripts.
454 The predicted gene sequences were the basis for an evaluation of the completeness of the
455 assembly and gene prediction by the program BUSCO v.3 [58]. We used this method with the
456 lineage dataset *Pezizomycotina.* The predicted genes were also used to create a phylogeny with
457 the online implementation of CVtree3, using kmer sizes of 6 and 7 as recommend for fungi [59].
458 We generated a second tree with the whole assemblies, estimating a distance matrix using the
459 andi software [60].
460 We predicted orthologs between the newly assembled genomes, the reference *Z. tritici* genome
461 and the reference *Cercospora beticola* genome, a related Dothideomycete, which we used as

462 outgroup to identify genes with orthologs restricted to the *Zymoseptoria* genus [21]. For this, we
463 used the software PoFF [35, 61] which takes into account synteny information in the analyses of
464 similarity inferred by the program Proteinortho [35].These orthogroups were used to visualize
465 synteny between genomes using Circos [62].

466 The whole-genome assemblies were used to create a matrix distance with the software andi,
467 from which we generated a tree [60]. A second tree was generated from the gene prediction with
468 the online implementation of CVtree3 [59].

469 **Functional annotations**

470 We used several tools to predict the putative functions for the gene models. First, we used the
471 eggnog-mapper which provide COG, GO and KEGG annotations [63]. The online resource
472 dbCAN2 was run to identify carbohydrate-active enzymes (CAZymes) [64]. Finally, for each
473 genome, we used Antismash v3 (fungal version) to detect biosynthetic gene clusters (Figure S1;
474 [65].
475 Additionally, we designed a pipeline to predict protein cell localization and to identify effector
476 candidates. The pipeline for effector prediction is outlined in Figure S1 and includes the software
477 DeepLoc [66], SignalP [67], TargetP [68], phobius [69] and TMHMM [70, 71], which predict the
478 cellular location, the peptide signals and whether proteins are transmembrane. Effectors were
479 identified with EffectorP v2 which uses both a new machine learning approach and more
480 complete databases to improve effector prediction compared to the previous version [19]. The
481 pipeline also includes software which are specifically targeted to annotate plant pathogenic
482 functions, namely the program ApoplastP [72] and LOCALIZER [73]. We wrote wrappers scripts,
483 which run the software and create consensus between the different prediction tools providing
484 one command line from the user. These scripts are available at
485 https://gitlab.gwdg.de/alice.feurtey/genome_architecture_zymoseptoria. Briefly, we gathered
486 outputs of several software to predict the cellular location, transmembrane domain and secretion
487 and created a consensus based on the different output to prevent the pitfalls of any one of these
488 methods. From this consensus, we extracted the gene products predicted to be secreted and

489 without a transmembrane domain. The comparisons of genes functions repartition were done by
490 combining predictions of COG categories, secondary metabolite genes with pathogenicity-
491 related gene functional categories such as CAZymes and effector predictions.

492 **Gene expression analyses**

493 To update expression profiles on the new genome assemblies and new gene predictions of the
494 three *Z. tritici* isolates, we used previously generated RNA-seq data from *in planta* and *in vitro*
495 growth [30, 32]. The *in planta* RNAseq data was obtained from infected leaves at four different
496 stages corresponding to early and late biotrophic (stage A and stage B) and necrotrophic (stage
497 C and stage D) stages of the three *Z. tritici* isolates [30]. Strand-specific RNA-libraries were
498 sequenced using Illumina HISeq2500, with 100pb single-end reads for a total read number
499 ranging from 89.5 to 147.5 million reads per sample. This data was previously analyzed [30],
500 using gene predictions generated from an Illumina-based assembly [22]. The reads were here
501 mapped on the new assemblies of Zt5 and Zt10 and the reference genome of IPO323 after
502 trimming. We used the DESeq2 R package to determine differential gene expression during *in*
503 *planta* infection, considering only two infection stages; biotrophic and necrotrophic [74]. Gene
504 expression was assessed as Transcript per Million (TPM). Briefly, TPM is calculated by
505 normalizing read counts with coding region length resulting in the number of reads per kilobase
506 (RPK). RPK total counts per sample are then divided by 1 million to generate a “per million”
507 scaling factor. We calculated the coding region length of each gene with GenomicFeatures R
508 package using the function called “exonsBy” [75]. For gene expression analyses, we further
509 filtered our gene predictions to remove any predicted transposases and other TE-related
510 annotations based on the Eggnog mapper annotations.

511 **ChIP-sequencing and data analysis**

512 *Z. ardabiliae* (Za17) and *Z. pseudotritici* (Zp13) cells were grown in liquid YMS medium for two
513 days at 18°C until an OD_{600} of ~ 1 was reached. Chromatin immunoprecipitation and library
514 preparation were performed as previously described [76]. We sequenced two biological and two
515 technical replicates per isolate and used antibodies against the euchromatin histone mark

H3K4me2 (#07–030, Merck Millipore). Sequencing was performed at the OSU Center for Genome Research and Biocomputing (Oregon State University, Corvallis, USA) on an Illumina HiSeq2000 to obtain 50-nt reads. The data was quality-filtered using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), mapping was performed using bowtie2 [77] and peaks were called using HOMER [78]. Peaks were called individually for each replicate, but only peaks that were detected in all replicates were considered and merged for further analysis. Merging of peaks and genome wide sequence coverage with enriched regions was assessed using bedtools [46].

Data availability

The assembled genomes can be found at 10.5281/zenodo.3568213. The gene annotations are deposited at 10.5281/zenodo.3568213. The functional annotation pipeline, additional scripts and command lines used to create the results presented in this manuscript can be found at https://gitlab.gwdg.de/alice.feurtey/genome_architecture_zymoseptoria.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the data supporting the findings of this study are openly available at 10.5281/zenodo.3568213.

Competing interests

The authors declare that they have no competing interests.

Funding and Acknowledgements

CL is funded by the Institut national de la recherche agronomique (INRA) in the framework of a “Contrat Jeune Scientifique” and by the Labex ARBRE (Lab of Excellence ARBRE). Genome research in the group of EHS is funded by the DFG Priority Program SPP1819 and the

Canadian Institute for Advanced Research (CIFAR). AF is funded by by the DFG Priority Program SPP1819.

Authors' contributions

AF, CL and MM performed data and results analyses. EHS, AF and CL contributed to the design and implementation of the research. AF, CL, CE and DC performed genome assemblies. MH, JH, MM, MF and KS performed the experimental procedures. All authors contributed to the writing of the manuscript.

References

1. Cook DE, Mesarich CH, Thomma BPHJ. Understanding plant immunity as a surveillance system to detect invasion. Annu Rev Phytopathol. 2015;53:541–63.

2. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, et al. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nat Commun. 2011;2:202.

3. Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 2015;35:57–65. doi:10.1016/j.gde.2015.09.001.

4. Todd RT, Wikoff TD, Forche A, Selmecki A. Genome plasticity in Candida albicans is driven by long repeat sequences. Elife. 2019;8. doi:10.7554/eLife.45954.

5. Mehrabi R, Mirzadi Gohari A, Kema GHJ. Karyotype Variability in Plant-Pathogenic Fungi. Annu Rev Phytopathol. 2017;55:483–503. doi:10.1146/annurev-phyto-080615-095928.

6. Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, et al. Histone modifications rather than the novel regional centromeres of *Zymoseptoria tritici* distinguish core and accessory chromosomes. Epigenetics Chromatin. 2015;8:41. doi:10.1186/s13072-015-0033-5.

7. Fokkens L, Shahi S, Connolly LR, Stam R, Schmidt SM, Smith KM, et al. The multi-speed genome of Fusarium oxysporum reveals association of histone modifications with sequence divergence and footprints of past horizontal chromosome transfer events. bioRxiv. 2018;:465070.

569 8. Miao VP, Covert SF, Vanetrent HD. A Fungal Gene for Antibiotic Resistance on a
570 Dispensable (" B ") Chromosome. Science (80-). 1991;254. doi:10.1126/science.1763326.
571 9. Temporini E, VanEtten H. An analysis of the phylogenetic distribution of the pea pathogenicity
572 genes of Nectria haematococca MPVI supports the hypothesis of their origin by horizontal
573 transfer and uncovers a potentially new pathogen of garden pea: Neocosmospora boniensis.
574 Curr Genet. 2004;46:29–36. doi:10.1007/s00294-004-0506-8.
575 10. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-JM-J, Di Pietro A, et al.
576 Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature.
577 2010;464:367–73.
578 11. Balesdent M-H, Fudal I, Ollivier B, Bally P, Grandaubert J, Eber F, et al. The dispensable
579 chromosome of Leptosphaeria maculans shelters an effector gene conferring avirulence towards
580 Brassica rapa. New Phytol. 2013;198:887–98.
581 12. van der Does HC, Rep M. Adaptation to the Host Environment by Plant-Pathogenic Fungi.
582 Annu Rev Phytopathol. 2017;55:427–50.
583 13. Lorrain C, Petre B, Duplessis S. Show me the way: rust effector targets in heterologous plant
584 systems. Curr Opin Microbiol. 2018;46:19–25.
585 14. Toruño TY, Stergiopoulos I, Coaker G. Plant-Pathogen Effectors: Cellular Probes Interfering
586 with Plant Defenses in Spatial and Temporal Manners. Annu Rev Phytopathol. 2016;54:419–41.
587 doi:10.1146/annurev-phyto-080615-100204.
588 15. Zhao Z, Liu H, Wang C, Xu J-R. Comparative analysis of fungal genomes reveals different
589 plant cell wall degrading capacity in fungi. BMC Genomics. 2013;14:274. doi:10.1186/1471-
590 2164-14-274.
591 16. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, et al. Fungal Effectors and
592 Plant Susceptibility. Annu Rev Plant Biol. 2015;66:513–45. doi:10.1146/annurev-arplant-043014-
593 114623.
594 17. Snelders NC, Kettles GJ, Rudd JJ, Thomma BPHJ. Plant pathogen effector proteins as
595 manipulators of host microbiomes? Mol Plant Pathol. 2018;19:257–9.
596 18. Shi-Kunne X, Jové R de P, Depotter JRL, Ebert MK, Seidl MF, Thomma BPHJ. In silico

597 prediction and characterisation of secondary metabolite clusters in the plant pathogenic fungus
598 Verticillium dahliae. FEMS Microbiol Lett. 2019;366.
599 19. Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. Improved prediction of
600 fungal effector proteins from secretomes with EffectorP 2.0. Mol Plant Pathol. 2018;19:2094–
601 110. doi:10.1111/mpp.12682.
602 20. Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. Origin and domestication of the
603 fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. Mol Biol Evol.
604 2007;24:398–411.
605 21. Stukenbrock EH, Quaedvlieg W, Javan-Nikhah M, Zala M, Crous PW, McDonald BA.
606 *Zymoseptoria ardabiliae* and *Z. pseudotritici*, two progenitor species of the septoria tritici leaf
607 blotch fungus *Z. tritici* (synonym: *Mycosphaerella graminicola*). Mycologia. 2012;104:1397–407.
608 doi:10.3852/11-374.
609 22. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq-Based Gene Annotation and
610 Comparative Genomics of Four Fungal Grass Pathogens in the Genus *Zymoseptoria* Identify
611 Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3 (Bethesda).
612 2015;5:1323–33. doi:10.1534/g3.115.017731.
613 23. Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished
614 Genome of the Fungal Wheat Pathogen Mycosphaerella graminicola Reveals Dispensome
615 Structure, Chromosome Plasticity, and Stealth Pathogenesis. PLoS Genet. 2011;7:e1002070.
616 doi:10.1371/journal.pgen.1002070.
617 24. Habig M, Quade J, Stukenbrock EH. Forward Genetics Approach Reveals Host Genotype-
618 Dependent Importance of Accessory Chromosomes in the Fungal Wheat Pathogen
619 Zymoseptoria tritici. MBio. 2017;8:e01919-17.
620 25. Habig M, Kema G, Holtgrewe Stukenbrock E. Meiotic drive of female-inherited
621 supernumerary chromosomes in a pathogenic fungus. Elife. 2018.
622 26. Möller M, Habig M, Freitag M, Stukenbrock EH. Extraordinary Genome Instability and
623 Widespread Chromosome Rearrangements During Vegetative Growth. Genetics.
624 2018;210:517–29. doi:10.1534/genetics.118.301050.

625 27. Fouché S, Plissonneau C, McDonald BA, Croll D. Meiosis Leads to Pervasive Copy-Number
626 Variation and Distorted Inheritance of Accessory Chromosomes of the Wheat Pathogen
627 Zymoseptoria tritici. Genome Biol Evol. 2018;10:1416–29. doi:10.1093/gbe/evy100.
628 28. Plissonneau C, Hartmann FE, Croll D. Pangenome analyses of the wheat pathogen
629 *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol.
630 2018;16:5. doi:10.1186/s12915-017-0457-4.
631 29. Hartmann FE, McDonald BA, Croll D. Genome-wide evidence for divergent selection
632 between populations of a major agricultural pathogen. Mol Ecol. 2018;27:2725–41.
633 30. Haueisen J, Möller M, Eschenbrenner CJ, Grandaubert J, Seybold H, Adamiak H, et al.
634 Highly flexible infection programs in a specialized wheat pathogen. Ecol Evol. 2018.
635 doi:10.1002/ece3.4724.
636 31. Kellner R, Bhattacharyya A, Poppe S, Hsu TY, Brem RB, Stukenbrock EH. Expression
637 profiling of the wheat pathogen *Zymoseptoria tritici* reveals genomic patterns of transcription and
638 host-specific regulatory programs. Genome Biol Evol. 2014;6:1353–65. doi:10.1093/gbe/evu101.
639 32. Möller M, Schotanus K, Soyer JL, Haueisen J, Happ K, Stralucke M, et al. Destabilization of
640 chromosome structure by histone H3 lysine 27 methylation. PLOS Genet. 2019;15:e1008093.
641 doi:10.1371/journal.pgen.1008093.
642 33. Plissonneau C, Stürchler A, Croll D. The Evolution of Orphan Regions in Genomes of a
643 Fungal Pathogen of Wheat. MBio. 2016;7:e01231-16.
644 34. Vaghefi N, Kikkert JR, Bolton MD, Hanson LE, Secor GA, Pethybridge SJ. De novo genome
645 assembly of Cercospora beticola for microsatellite marker development and validation. Fungal
646 Ecol. 2017;26:125–34.
647 35. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: Detection
648 of (Co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011;12:124. doi:10.1186/1471-
649 2105-12-124.
650 36. Lorrain C, Hecker A, Duplessis S. Effector-Mining in the Poplar Rust Fungus Melampsora
651 larici-populina Secretome . Frontiers in Plant Science . 2015;6:1051.
652 37. Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. A 19-isolate reference-quality

653 global pangenome for the fungal wheat pathogen Zymoseptoria tritici. bioRxiv. 2019;:803098.
654 doi:10.1101/803098.
655 38. Feurtey A, Stevens DM, Stephan W, Stukenbrock EH. Interspecific Gene Exchange
656 Introduces High Genetic Variability in Crop Pathogen. Genome Biol Evol. 2019;11:3095–105.
657 doi:10.1093/gbe/evz224.
658 39. Wu B, Macielog AI, Hao W. Origin and Spread of Spliceosomal Introns: Insights from the
659 Fungal Clade *Zymoseptoria*. Genome Biol Evol. 2017;9:2658–67. doi:10.1093/gbe/evx211.
660 40. Schulze-Lefert P, Panstruga R. A molecular evolutionary concept connecting nonhost
661 resistance, pathogen host range, and pathogen speciation. Trends Plant Sci. 2011;16:117–25.
662 41. Valero-Jiménez CA, Veloso J, Staats M, van Kan JAL. Comparative genomics of plant
663 pathogenic Botrytis species with distinct host specificity. BMC Genomics. 2019;20:203.
664 42. Thines M. An evolutionary framework for host shifts – jumping ships for survival. New Phytol.
665 2019;:605–17.
666 43. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. A modified protocol
667 for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc.
668 2006;1:2320–5.
669 44. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
670 assemblies. Bioinformatics. 2013;29:1072–5. doi:10.1093/bioinformatics/btt086.
671 45. Fulnecková J, Sevcíková T, Fajkus J, Lukesová A, Lukes M, Vlcek C, et al. A broad
672 phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biol
673 Evol. 2013;5:468–83. doi:10.1093/gbe/evt019.
674 46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
675 Bioinformatics. 2010;26:841–2. doi:10.1093/bioinformatics/btq033.
676 47. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory
677 requirements. Nat Methods. 2015;12:357–60. doi:10.1038/nmeth.3317.
678 48. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element
679 Diversification in De Novo Annotation Approaches. PLoS One. 2011;6:e16526.
680 doi:10.1371/journal.pone.0016526.

681 49. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al.
682 Combined Evidence Annotation of Transposable Elements in Genome Sequences. PLoS
683 Comput Biol. 2005;1:e22. doi:10.1371/journal.pcbi.0010022.
684 50. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel
685 fungal genomes using an ab initio algorithm with unsupervised training. Genome Res.
686 2008;18:1979–90. doi:10.1101/gr.081612.108.
687 51. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
688 Bioinformatics. 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.
689 52. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-
690 Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. Bioinformatics.
691 2016;32:767–9. doi:10.1093/bioinformatics/btv661.
692 53. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped
693 cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
694 doi:10.1093/bioinformatics/btn013.
695 54. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into
696 automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42:e119–e119.
697 doi:10.1093/nar/gku557.
698 55. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
699 transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol.
700 2011;29:644–52. doi:10.1038/nbt.1883.
701 56. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene
702 structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.
703 Genome Biol. 2008;9:R7. doi:10.1186/gb-2008-9-1-r7.
704 57. Gremme G, Steinbiss S, Kurtz S. GenomeTools: A Comprehensive Software Library for
705 Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans Comput Biol
706 Bioinforma. 2013;10:645–56. doi:10.1109/TCBB.2013.68.
707 58. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al.
708 BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol

709 Biol Evol. 2018;35:543–8. doi:10.1093/molbev/msx319.

710 59. Zuo G, Hao B. CVTree3 Web Server for Whole-genome-based and Alignment-free

711 Prokaryotic Phylogeny and Taxonomy. Genomics Proteomics Bioinformatics. 2015;13:321–31.

712 doi:10.1016/J.GPB.2015.08.004.

713 60. Haubold B, Klötzl F, Pfaffelhuber P. andi: Fast and accurate estimation of evolutionary

714 distances between closely related genomes. Bioinformatics. 2015;31:1169–75.

715 doi:10.1093/bioinformatics/btu815.

716 61. Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, et al.

717 Orthology detection combining clustering and synteny for very large datasets. PLoS One.

718 2014;9:e105015. doi:10.1371/journal.pone.0105015.

719 62. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an

720 information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

721 63. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast

722 Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol

723 Biol Evol. 2017;34:2115–22. doi:10.1093/molbev/msx148.

724 64. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for

725 automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46:W95–101.

726 doi:10.1093/nar/gky418.

727 65. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a

728 comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids

729 Res. 2015;43:W237–43. doi:10.1093/nar/gkv437.

730 66. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc:

731 prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33:3387–

732 95. doi:10.1093/bioinformatics/btx431.

733 67. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides

734 from transmembrane regions. Nat Methods. 2011;8:785–6. doi:10.1038/nmeth.1701.

735 68. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting Subcellular Localization of

736 Proteins Based on their N-terminal Amino Acid Sequence. J Mol Biol. 2000;300:1005–16.

737 doi:10.1006/jmbi.2000.3903.

738 69. Käll L, Krogh A, Sonnhammer EL. A Combined Transmembrane Topology and Signal

739 Peptide Prediction Method. J Mol Biol. 2004;338:1027–36. doi:10.1016/j.jmb.2004.03.016.

740 70. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein

741 topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J

742 Mol Biol. 2001;305:567–80. doi:10.1006/jmbi.2000.4315.

743 71. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting

744 transmembrane helices in protein sequences. Proceedings Int Conf Intell Syst Mol Biol.

745 1998;6:175–82. http://www.ncbi.nlm.nih.gov/pubmed/9783223. Accessed 26 Jun 2019.

746 72. Sperschneider J, Dodds PN, Singh KB, Taylor JM. ApoplastP□: prediction of effectors and

747 plant proteins in the apoplast using machine learning. New Phytol. 2018;217:1764–78.

748 doi:10.1111/nph.14946.

749 73. Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, et al.

750 LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant

751 cell. Sci Rep. 2017;7:44598. doi:10.1038/srep44598.

752 74. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-

753 seq data with DESeq2. Genome Biol. 2014;15:550. doi:10.1186/s13059-014-0550-8.

754 75. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for

755 Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013;9:e1003118.

756 doi:10.1371/journal.pcbi.1003118.

757 76. Soyer JL, Möller M, Schotanus K, Connolly LR, Galazka JM, Freitag M, et al. Chromatin

758 analyses of Zymoseptoria tritici: Methods for chromatin immunoprecipitation followed by high-

759 throughput sequencing (ChIP-seq). Fungal Genet Biol. 2015;79:63–70.

760 doi:10.1016/j.fgb.2015.03.006.

761 77. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.

762 2012;9:357–9. doi:10.1038/nmeth.1923.

763 78. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of

764 lineage-determining transcription factors prime cis-regulatory elements required for macrophage

765 and B cell identities. Mol Cell. 2010;38:576–89. doi:10.1016/j.molcel.2010.05.004.

766

767 Table 1: Metrics of genome assemblies and annotation

Species	***Zymoseptoria tritici***		***Zymoseptoria pseudotritici***	***Zymoseptoria brevis***	***Zymoseptoria ardabiliae***	***Zymoseptoria passerinii***
Isolate	Zt05	Zt10	Zp13	Zb87	Za17	Zpa63
Origin	Denmark	Iran, Ilam province	Iran, Ardabil province	Iran	Iran, Ardabil province	USA
Host	*Triticum aestivum*	*Triticum aestivum*	*Dactylis glomerata*	*Phalaris paradoxa*	*Lolium perenne*	*Hordeum vulgare*
Year (of isolation)	2004	2001	2004		2004	
Contig number	30	19	42	29	50	103
Total length (bp)	41240984	39248105	40312446	41586671	38100668	41398787
Mean contig size (bp)	1374699	2065690	59820	1434023	762013	401930
N50	2454671	2925395	2115121	2744794	1156695	737698
L50	6	5	7	7	11	18
Contigs with telomeric repeats on both ends	12	19	5	9	0	0
Number of genes	12386	11991	11661	11480	11463	10528
Repeat content (%)	19.9	16.5	20.8	29.2	18.2	31.4

Figure Legends

Figure 1: Whole-genome phylogeny of *Zymoseptoria* spp. and basic statistics for the assemblies and gene predictions. Tree based on the distance matrix generated from the whole-genome assemblies using the software andi (Haubold et al. 2015). The bar plots represent the number of genes coding for secreted proteins (pink) and non-secreted proteins (grey) for each genome.

Figure 2: Intra- and inter-species synteny conservation in *Zymoseptoria* genus. A) Intra-species synteny between the reference genome of *Z. tritici* and the genome of the Iranian *Z. tritici* isolate Zt10. Each color represents a different chromosome as defined in the reference *Z. tritici* genome, except for accessory chromosomes, which are in grey. The links represent orthologous genes. B) Inter-species synteny between the reference genome of *Z. tritici* and the genome of *Z. brevis*. The arrows represent the large-scale inversions identified between the genomes of these two species.

Figure 3: Genome architecture of the reference genome *Z. tritici* IPO323 (A) and *Z. pseudotritici* Zp13 (B). The segments constituting the first circle represents the chromosomes of IPO323 (A) and contigs of Zp13 (B) ordered according to chromosome numbers of the reference genome. Tracks from the outside to the inside are heatmaps representing respectively: gene density along chromosomes/contigs; gene expression in vitro (TPM); H3K4me2 levels in vitro and species-specific gene density. The arrows indicate the location of the region on chromosome 7 (and the corresponding syntenic region in Z.pseudotritici) displaying accessory-like genomic and regulatory hallmarks.

Figure 4: Orthogroups and functional gene categories in *Zymoseptoria* spp. genomes. A) Orthogroups shared by the reference *Z. tritici* genome, our new *Zymoseptoria* assemblies and the outgroup genome of *C. beticola*. Only intersects higher than 100 are displayed on the upset plot. The doughnut plot summarizes the number of orthologs grouped by larger categories: specific to some isolates, to a species or shared by all. The colored bars under the upset plot link each intersect to its corresponding category in the doughnut plot. B) Venn diagram representing the genes shared by the three isolates of *Z. tritici*. C) The only gene category found to be overrepresented in any of the specificity categories - other than unknown function genes - are effectors. Effectors genes are overrepresented in the genus-specific genes and in *Z. ardabiliae* specific genes (*** represent Fisher exact test p-value < 0.05).

Figure 5: Expression of genes belonging to different specificity levels in the *Zymoseptoria* pangenome. The boxplots represent the expression levels in both biotrophic and necrotrophic phase in transcript per million (TPM) for A) the whole transcriptome of *Z. tritici* isolates and B) in planta differentially expressed genes identified by DESeq2. Comparisons are performed by Krustal-Wallis test, different letters represent p-value < 0.05.

Supplementary Tables

Table S1: Summary statistics from gene predictions generated in this study.

Table S2: Summary of gene expression in *Z. tritici* isolates during infection stages

Table S3: Enrichment of species-specific and isolate-specific genes per chromosome/contig

Supplementary Figure Legends

Figure S1: Plant-associated genes compartmentalization along the chromosomes. The first track represents core (dark grey) and accessory (light grey) chromosomes/contigs. Fully assembled contigs are marked by the presence of telomeric repeats (orange). Circles from outside to inside represent the position of: effector genes (blue), biosynthetic gene clusters (BGC, green) and CAZymes (yellow), respectively.

Figure S2: A) Intra-species synteny between the reference genome of *Z. tritici* and the genome of the *Z. tritici* isolate Zt05. Each color represents a different chromosome as based on the reference *Z. tritici* genome, except for accessory chromosomes, which are in grey. The connecting lines represent orthologs between each genome. The track between the chromosomes and connecting lines are predicted effectors. B) Inter-species synteny between reference genome of *Z. tritici* and *Z. pseudotritici*. The arrows represent the large-scale inversions identified between the genomes of these two species.

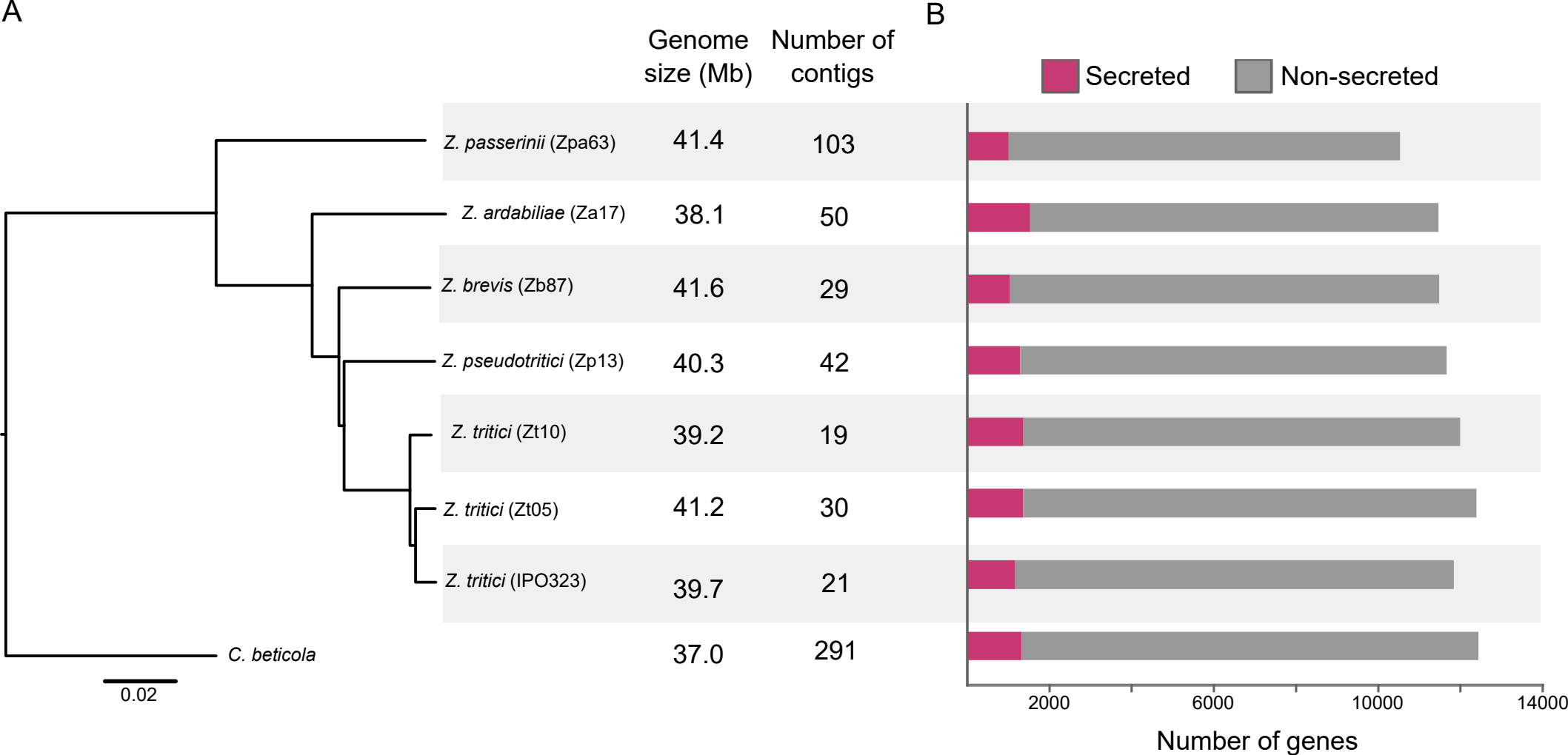
Figure S3: Inter-species synteny between genomes of *Z. pseudotritici* (dark grey) and *Z. brevis* (blue). The arrows indicate the contigs affected by the large-scale inversions identified between *Z. tritici* and both *Z.pseudotritici* and *Z. brevis*.

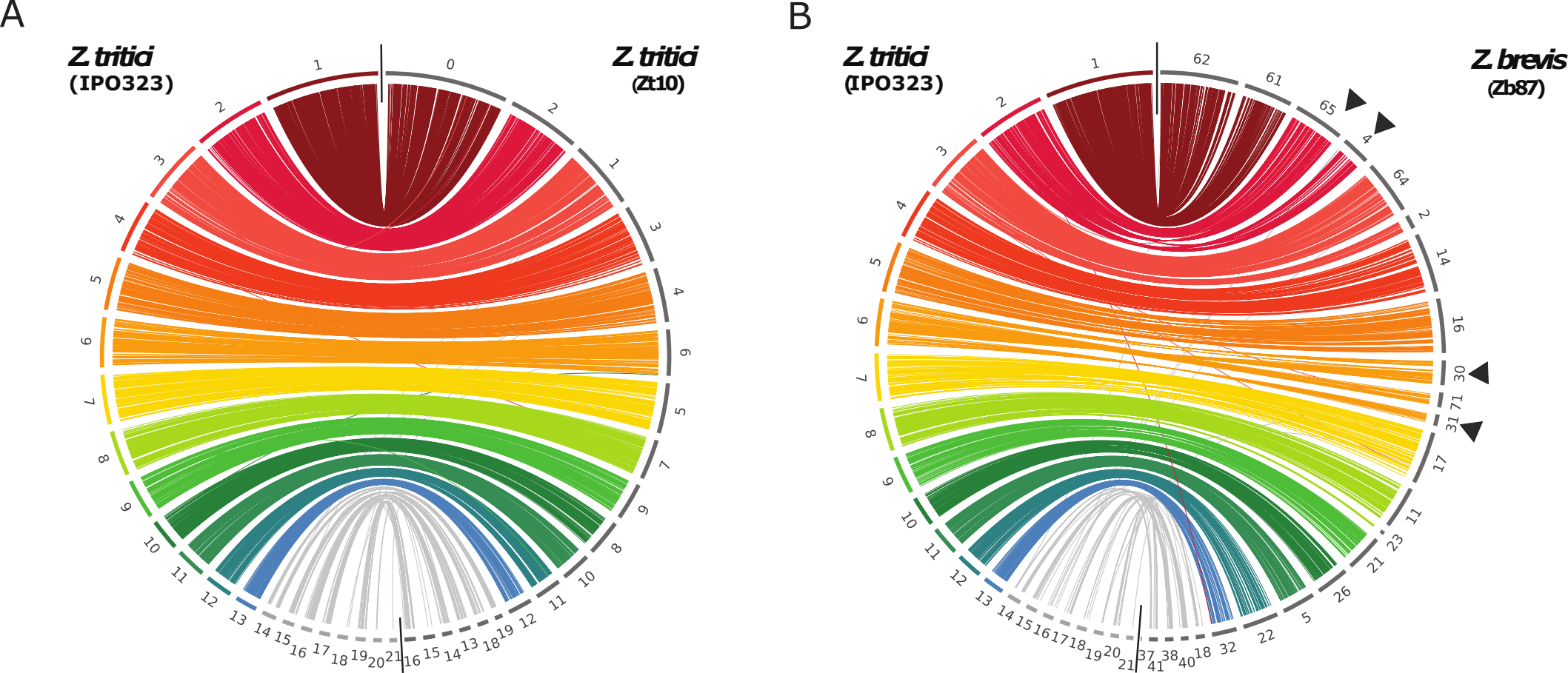
Figure S4: Genome architecture in A) *Z. tritici* Zt05, B) *Z. tritici* Zt10. C) *Z. brevis* Zb87 and D) *Z. ardabiliae* Za17. Circles from the outside to the inside represent respectively: gene density along chromosomes/contigs; gene expression in vitro (TPM); H3K4me2 distribution in vitro (only for Za17) and species-specific gene density

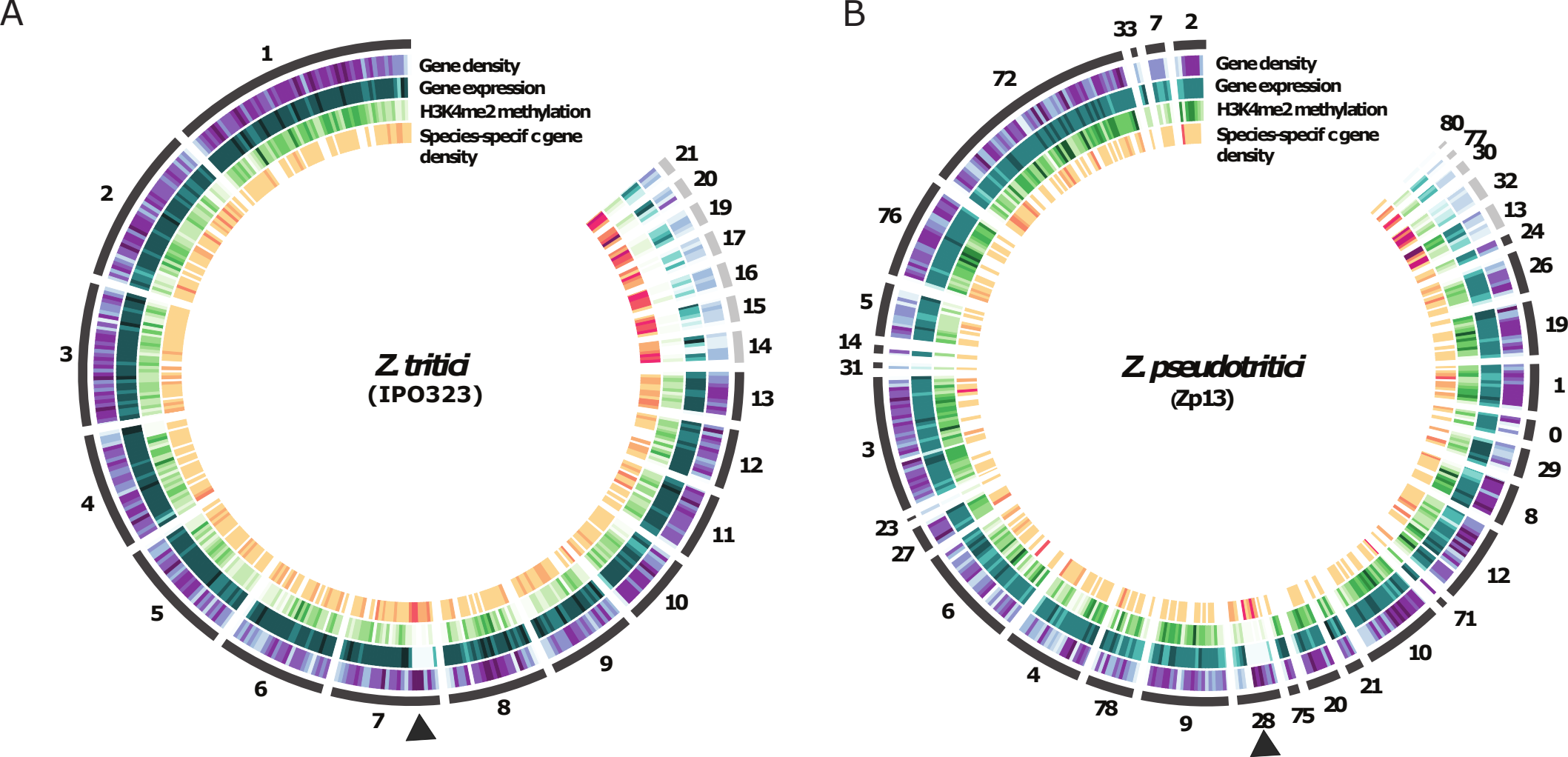
Figure S5: Simplified diagram of the pipeline used to predict the functions and subcellular localization of gene model products.

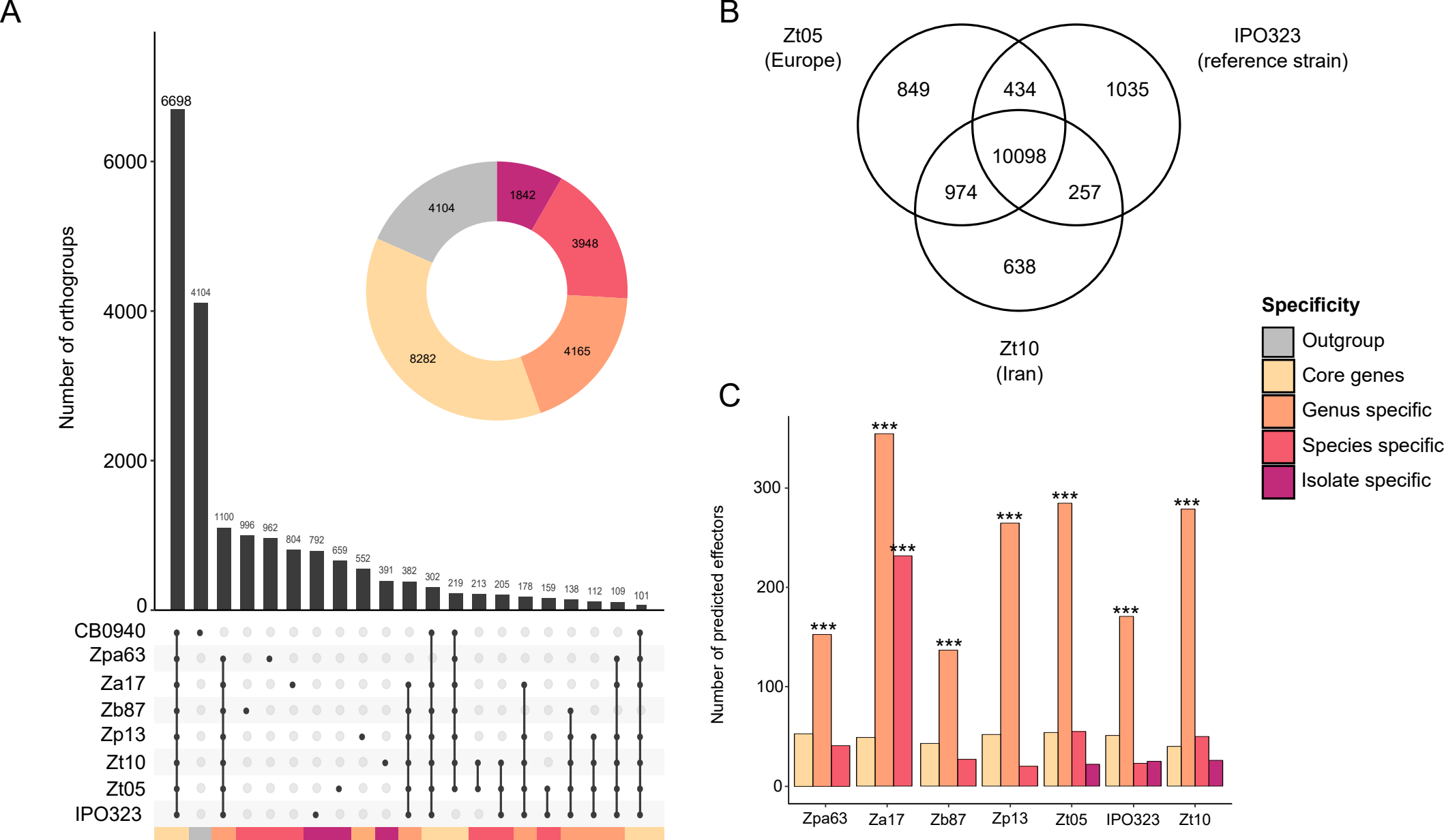
Figure S6: Functional gene categories in *Zymoseptoria spp*. genomes. A) The number of genes in Eggmapper COG categories in addition to Effectors and CAZymes. B) Pathogenicity-related genes of interest: secreted proteins, effectors, secreted CAZymes and non-secreted CAZymes. C) Subcellular localization of predicted gene products.

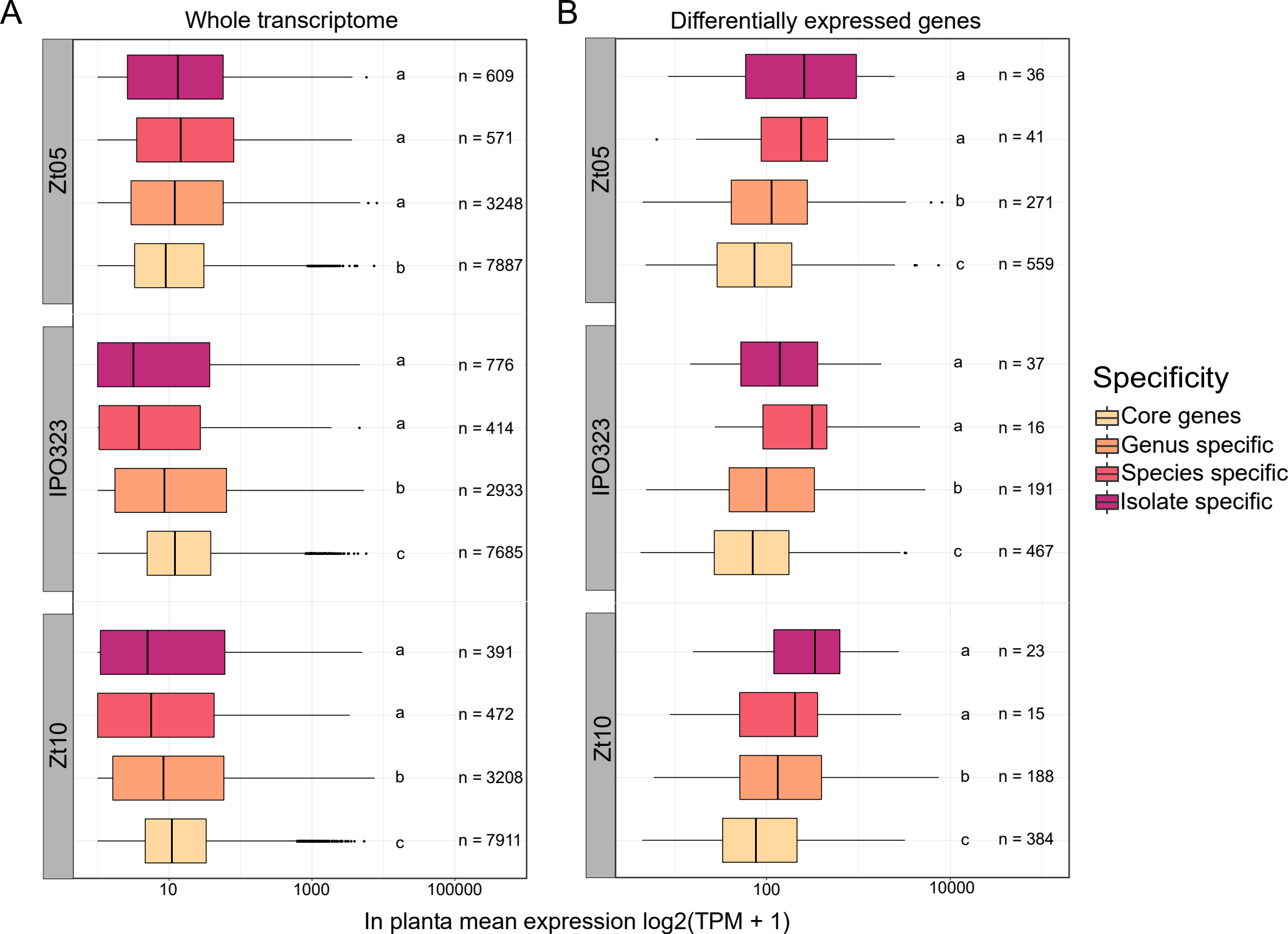
Figure S7: Principal component analysis of DESeq2 rlog transformed expression data.


A
B
Genome size (Mb)
Number of contigs
Secreted
Non-secreted
Z. passerinii (Zpa63)
41.4
103
Z. ardabiliae (Za17)
38.1
50
Z. brevis (Zb87)
41.6
29
Z. pseudotritici (Zp13)
40.3
42
Z. tritici (Zt10)
39.2
19
Z. tritici (Zt05)
41.2
30
Z. tritici (IPO323)
39.7
21
C. beticola
37.0
291
0.02
2000
6000
10000
14000
Number of genes


A
Z tritici
(IPO323)
Z tritici
(Zt10)
B
Z tritici
(IPO323)
Z brevis
(Zb87)


A
Gene density
Gene expression
H3K4me2 methylation
Species-specif c gene density
Z. tritici
(IPO323)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
20
21
B
Gene density
Gene expression
H3K4me2 methylation
Species-specif c gene density
Z. pseudotritici
(Zp13)
33
7
2
72
76
5
14
31
3
23
27
6
4
78
9
28
75
20
21
10
71
12
8
29
0
1
19
26
24
13
32
30
77
80


A
Number of orthogroups
6000
4000
2000
0
6698
4104
1100
996
962
804
792
659
552
391
382
302
219
213
205
178
159
138
112
109
101
CB0940
Zpa63
Za17
Zb87
Zp13
Zt10
Zt05
IPO323
4104
1842
3948
8282
4165
B
Zt05
(Europe)
IPO323
(reference strain)
849
434
1035
10098
974
257
638
Zt10
(Iran)
C
Number of predicted effectors
300
200
100
0
***
Zpa63
Za17
Zb87
Zp13
Zt05
IPO323
Zt10
Specificity
Outgroup
Core genes
Genus specific
Species specific
Isolate specific


A
Whole transcriptome
B
Differentially expressed genes
Zt05
IPO323
Zt10
a
b
c
n = 609
n = 571
n = 3248
n = 7887
n = 776
n = 414
n = 2933
n = 7685
n = 391
n = 472
n = 3208
n = 7911
n = 36
n = 41
n = 271
n = 559
n = 37
n = 16
n = 191
n = 467
n = 23
n = 15
n = 188
n = 384
10
1000
100000
100
10000
In planta mean expression log2(TPM + 1)
Specificity
Core genes
Genus specific
Species specific
Isolate specific