A large accessory genome, high recombination rates, and selection of secondary metabolite genes help maintain global distribution and broad host range of the fungal plant pathogen *Claviceps purpurea*

Stephen A. Wyka[1], Stephen J. Mondo[1,2], Miao Liu[3], Vamsi Nalam[1], Kirk D. Broders[4*]

[1] Department of Agricultural Biology, Colorado State University, Fort Collins, Colorado, United States of America.

[2] United States Department of Energy Joint Genome Institute, Berkeley, California, United States of America.

[3] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada.

[4] Smithsonian Tropical Research Institute, Apartado Panamá, República de Panamá.

[*] Corresponding authors: Kirk D. Broders, email: BrodersK@si.edu

Abstract:

Pangenome analyses are increasingly being utilized to study the evolution of eukaryotic organisms, which is often governed by variable gene content. While pangenomes can provide insight into polymorphic gene content, inferences about the ecological and adaptive potential of such organisms also need to be accompanied by additional supportive genomic analyses. In this study we constructed a pangenome of *Claviceps purpurea* from 24 genomes and examined the positive selection and recombination landscape of an economically important fungal organism for pharmacology and agricultural research. Together, these analyses revealed that *C. purpurea* has a relatively large accessory genome (~ 38%) that is likely maintained by high recombination rates ($\rho = 0.044$) and transposon mediated gene duplication. However, due to observations of relatively low transposable element (TE) content (8.8%) and a lack of variability in genome sizes, prolific TE expansion is likely controlled by these high recombination rates, which may additionally be influencing the overall trend of purifying selection across the genome. Despite this trend, we observed a strong positive selection pressure on secondary metabolite genes, particularly within the ergoline biosynthetic cluster where we also revealed that the *lpsA1* and *lpsA2* genes were the result of a recombination event. These results indicate that secondary metabolites are primary factors affecting the diversification of the species into new ecological niches and help maintain its global distribution and broad host range. These results showcase the use of selection and recombination landscapes to identify mechanisms contributing to pangenome structure and primary factors influencing the evolution of an organism.

Keywords:

Pangenome, positive selection, adaptive evolution, ergot

Author Summary:

The use of genomic data to better understand the lifestyle of a pathogen and its relationship with its host has expanded our ability to investigate the evolutionary history of these organisms. This in turn has allowed us to decipher and understand the ambiguity surrounding the true nature of the fungal plant pathogen *Claviceps purpurea*. By combining three different types of broad genomic analyses we identified primary factors affecting the evolution and adaptive potential of this pathogen; particularly a large accessory genome, high recombination rates, and positive selection of genes associated with stress tolerance. These factors likely contribute to the pathogen's global distribution and broad host range. Furthermore, these findings will influence the direction of future research into optimal control methods.

Introduction:

Pangenomes can provide useful insight into a species distribution and lifestyle through examination of gene functional diversity, abundance, and distribution into core and accessory genomes. These variations often provide fitness advantages and promote adaptive evolution of the organism (Araki *et al.* 2006; Hartmann *et al.* 2018; Brynildsrud *et al.* 2019). In prokaryotes the existence of more open pangenomes (large accessory) has been suggested to be the result of adaptive evolution that allows organisms, with large effective population sizes, to migrate into new ecological niches (McInerney *et al.* 2017). Whereas closed pangenomes (larger core) are found to be associated with more obligate and specialized organisms (McInerney *et al.* 2017). Similar results have been identified in fungal species, where a range of saprotrophic to opportunistic yeasts were found to have accessory genomes representing ~ 9 – 19% of the genes

(McCarthy and Fitzpatrick 2019), while *Zymoseptoria tritici*, a global wheat pathogen, has 40% of genes in the accessory genome (Badet *et al.* 2020). This increase in the *Z. tritici* accessory genome reflects the global distribution of this pathogen that must continuously adapt to overcome new host resistances and multiple cycles of annual fungicide applications (Sánchez-Vallet *et al.* 2018; Badet *et al.* 2020). While the identification of pangenome sizes provide valuable knowledge of polymorphic gene content, which can be used to infer the lifestyle of the species (McInerney *et al.* 2017), a combination of pangenomic and alternative genomic analyses provide a deeper understanding of the primary factors that are contributing to pangenome structure and the adaptive trajectory of the organism.

Claviceps purpurea is a biotrophic ascomycete plant pathogen that has a specialized ovarian-specific non-systemic lifestyle with its grass hosts (Píchová *et al.* 2018). Despite the specialized infection pattern, *C. purpurea* has a broad host range of ~ 400 grass species across 8 grass tribes, including economically important cereal crops such as wheat, barley, and rye and has a global distribution (Píchová *et al.* 2018). However, the mechanisms that underlie the evolutionary success of this species is still understudied. Unlike other pathogens of cereal crops, researchers have been unsuccessful in identifying qualitative resistance (R) genes in crop or wild grass varieties (Menzies and Turkington 2015; Menzies *et al.* 2017; Gordon *et al.* 2020). Menzies *et al.* (2017) did note the potential for a complex virulence and host susceptibility relationship of *C. purpurea* on durum and hexaploid wheat varieties, however, virulence was determined if sclerotia weighed > 81 mg; indicating that *C. purpurea* is able to initiate its biotrophic interaction but might be arrested during the final stages of sclerotia development. During infection the fungus does not induce necrosis or hypersensitive response (host mediated cell death) in its host, instead it actively manages to maintain host cell viability to obtain

nutrients from living tissue through a complex cross-talk of fungal cytokinin production (Hinsch *et al.* 2015, 2016; Oeser *et al.* 2017; Kind *et al.* 2018a, 2018b). Furthermore, Wyka *et al.* (2020a) revealed evidence of tandem gene duplication occurring in genes often associated with pathogenicity or evasion of host defenses (effectors), which could implicate their role in the success of the species, however, the factors that were influencing these duplication events remain unclear.

Claviceps purpurea is also known for its diverse secondary metabolite profile of ergot alkaloids and pigments (Schardl *et al.* 2013; Tudzynski and Neubauer 2014; Neubauer *et al.* 2016; Flieger *et al.* 2019). Fungal secondary metabolites can play important roles in plant-host interactions as virulence factors but can also increase the fitness of the fungus through stress tolerance (Avalos and Carmen Limon, 2015; Píchová *et al.* 2018; Pusztahelyi *et al.* 2019). It was also recently postulated that the evolution of *C. purpurea* was associated with a host jump and subsequent adaptation and diversification to cooler, more open habitats (Píchová *et al.* 2018; Wyka *et al.* 2020a). In addition, likely due to the toxicity of ergot alkaloids, grass grazing mammals showed avoidance in grazing grass infected with *C. purpurea*, suggesting a potential for beneficial effects for the host plant (Wäli *et al.* 2013). This along with other evidence of neutral to positive effects of infection to host plants (Raybould *et al.* 1998; Fisher *et al.* 2007) suggest that *C. purpurea* is a conditional defensive mutualist (Wäli *et al.* 2013).

In this study, we implement a comprehensive population genomic analysis to gain a deeper understating of factors governing the evolution and adaptive potential of *C. purpurea*. Using 24 isolates, from six countries and three continents, we construct the pangenome and subsequently use single-copy core orthologs to identify genes under positive selection. Full genome alignments were further utilized to estimate population recombination rates and predict

recombination hotspots. We observed a large accessory genome likely maintained by a large effective population size and high recombination rates, which subsequently influence an overall trend of purifying selection and likely help defend against TE expansion. In addition, we observed that the *lpsA1* and *lpsA2* genes of the well-known ergoline biosynthetic cluster were likely the result of a recombination event.

Results:

Pangenome analysis:

We constructed a pangenome of *Claviceps purpurea* from 24 isolates representing a collection from three continents and six countries (Table 1). Taking advantage of plentiful isolates available from Canada, we sampled more heavily from different provinces and on different host plants. The principal component and phylogenetic analysis revealed substantial genetic variation among the samples. However, the genetic distances were not correlated with geographic distances, such as LM470 (Canada) and Clav04 (USA) grouping closer to isolates from Europe and the isolate from New Zealand (Additional File 1 Fig. S1). In addition, across Canada and USA, isolates from similar regions rarely clustered together and were often intermixed (Additional File 1 Fig. S1B). These results agree with the results from a multi-locus genotyping of extended samples from Canada and Midwestern USA (Liu *et al. unpublished data*). Previous reports (Wyka *et al.* 2020a) showed that *C. purpurea* isolates had similar genome size (30.5 Mb – 32.1 Mb), genomic GC content (51.6% - 51.8%), TE content (8.42% - 10.87%), gene content (8,394 – 8,824), and BUSCO completeness score (95.5% - 98.0%) (Table 1). The pangenome consisted of 205,354 genes which were assigned to 10,540 orthogroups. We

observed 6,558 (62.22%) orthogroups shared between all 24 isolates (core genome), of which 6,244 (59.2%) were single-copy gene clusters, while the remaining core orthogroups, 314 (3%), contained paralogs (2 – 8 paralogs per cluster). The accessory genome consisted of 3,982 (37.78%) orthogroups with 2,851 (27.05%) shared by at least two isolates (but not all) and 1,131 (10.73%) were lineage-specific (singletons) found in only one isolate (Fig. 1, Additional File 2 Table S1). Within the accessory genome (including lineage-specific orthogroups) we observed 592 (5.6%) orthogroups containing paralogs, with some isolates containing > 20 genes per cluster (Fig. 1C, Additional File 2 Table S1).

We utilized multiple gene functional categories to get a deeper understanding of how genes of different function were structured within the pangenome. As a proportion of orthogroups within each pangenome category (core, accessory, and singleton) we found that the core genome was significantly enriched in orthogroups that contained genes with conserved protein domains (conserved) (5,471; 84%), transmembrane domains (transmembrane) (1,038; 16%), peptidase and protease domains (MEROPs) (211, 3.2%), and orthogroups of carbohydrate-active enzymes (CAZys) (212, 3.2%) ($P < 0.01$, Fisher's exact test, Fig. 2A and 2E-G). Effector proteins play major roles in plant-microbe interactions, often conveying infection potential of the pathogen. A total of 257 predicted effector orthogroups were identified; 100 (38.9%) were core, 143 (55.6%) were accessory, and 14 (5.4%) were singletons. Predicted effectors and orthogroups coding for secreted proteins, which also contribute to host-pathogen interactions, were significantly enriched in the accessory genome (143, 5%; 218, 7.6%; respectively) ($P < 0.01$, Fisher's exact test, Fig. 2C and 2D). Although, the accessory and singleton genomes were largely composed of unclassified orthogroups (1791, 62.8%; 830, 73.4%; respectively) ($P < 0.01$, Fisher's exact test, Fig. 2H). Lastly, we observed that

orthogroups which contained secondary (2°) metabolite genes were similarly represented across all pangenome categories ($P > 0.05$, Fisher's exact test, Fig. 2B).

As expected, core orthogroups were found to be significantly enriched in general housekeeping and basic cellular functions and development such as protein and ATP binding, nucleus and membrane cellular components, and transmembrane transport, metabolic, and oxidation-reduction processes (Additional File 3 Table S2). Protein domains in core orthogroups were significantly enriched for several WD40-repeat domains, P-loop nucleoside triphosphate hydrolase (IPR027417), armadillo-type fold (IPR016024), and a major facilitator (PF07690) (Additional File 3 Table S2). When narrowing the focus to orthogroups with paralogs, core paralogous orthogroups were enriched in cytochrome P450 domains, and domains associated with trehalose activity (Additional File 3 Table S3). In contrast, the accessory genome was only found to be enriched in a fungal acid metalloendopeptidase domain (MER0001399) and the singleton genome had enrichment for a Tc5 transposase DNA-binding domain (PF03221) (Additional File 3 Table S2). Accessory paralogs were found to be enriched in several protein kinases, Myb-like domains, phosphotransferases, as well as DNA integration and a MULE transposase domain (Additional File 3 Table S3). Overall, our results revealed a large accessory pangenome enriched with genes associated with host-pathogen interactions and an abundance of orthogroups containing paralogs (8.6%), indicating the presence of proliferate gene duplication occurring within the species.

Positive selection landscape:

To further understand the evolution of genes within the pangenome we investigated the positive selection landscape on protein coding genes using 6,244 single-copy core orthologs to

177 compute the ratio of non-synonymous substitutions to synonymous substitutions (dN/dS). Ratios
178 of dN/dS (omega, ω) can provide information of evolutionary forces shaping an organism as
179 genes with ω > 1 may indicate positive or diversifying selection, ω = 1 may indicate neutral
180 evolution, and ω < 1 may indicate negative or purifying selection (Jeffares *et al.* 2015).

181 Overall, we saw low dN and dS values across all functional categories (Additional File 1
182 Fig. S3), corresponding to low ω ratios (Fig. 3). This suggests a general trend of purifying
183 selection within *C. purpurea*, although we did identify orthogroups with ω values > 1 (63, 1%),
184 of which 25 (40%) were unclassified (Fig. 3, Additional File 3 Table S4). Notable BLASTp
185 results showed that two conserved genes were related to transcription factors (OG0001193, ω =
186 1.13, related to subunits Tfc3; OG0004135, ω = 1.21, related to Cys6) and two were related to
187 DNA repair (OG0001034, ω = 1.05, related to mismatch repair PMS1; OG0004027, ω = 1.13,
188 related to XLF (XRCC4-like factor)) (Additional File 3 Table S5). The gene with the highest ω
189 was a transmembrane gene related to a bacteriophage N adsorption protein (OG0001093, ω =
190 9.79) (Additional File 3 Table S5). Overall, core unclassified genes showed the highest ω values
191 but were not significantly different than predicted effector genes ($P >> 0.05$, multi-test corrected
192 Mann-Whitney U Test, Fig. 3). In contrast, transmembrane, MEROPs, CAZys, and proteins with
193 conserved domains showed the lowest ω values, indicating that these genes are frequently
194 experiencing purifying selection.

195 While ω values, calculated across the entire gene, can provide useful insight on the
196 selective landscape of genes, positive selection and evolution occur at the codon triplet level and
197 can occur in genes where ω, across the entire gene, is < 1 (Goldman and Yang 1994). For this
198 reason, we utilized the CodeML algorithm (Yang 2007a) to more accurately and confidently
199 identify genes with signatures of positive selection. Our results revealed a total of 986 positively

selected genes (15.8%) that passed our stringent filtering (Fig. 4A). The majority were genes encoding conserved domains (557, 56.5%) followed by unclassified genes (192, 19.5%). While conserved genes made up the largest portion of genes under putative positive selection, unclassified genes showed the highest proportion of genes with positive selection signatures (26.8%) followed by secondary (2°) metabolite genes (21.5%) (Fig. 4B). We observed an enrichment of positively selected secondary metabolite genes that contained domains for polyketide synthases, several phosphopantetheines, as well as metabolic and catalytic GO terms ($P \leq 0.05$, Fisher's exact test, Fig. 4B, Additional File 3 Table S6). In addition, five genes in two known secondary metabolite clusters showed evidence of positive selective signatures; three genes (*easE* $\omega = 0.51$, *lpsB* $\omega = 0.34$, and *lps*C $\omega = 0.55$) in the well-known ergoline biosynthetic cluster (ergot alkaloids) (Schardl *et al.* 2013) and two genes (*tcpC* $\omega = 0.37$ and *tcp*P $\omega = 0.37$) in the epipolythiodiketopiperazine biosynthetic cluster (Dopstadt *et al.* 2016). Within these genes, positive selection was often observed in their AMP-binding and condensation domains but also occurred outside of the domain boundaries (Additional File 3 Table S7). Additionally, one of the three genes responsible for the biosynthesis of fungal cytokinins, a pisatin demethylase cytochrome P450 (Hinsch *et al.* 2015, 2016), had signatures of positive selection (OG0000984, $\omega = 0.19$) (Additional File 3 Table S4). Transmembrane genes saw enrichment of three multicopper oxidase domains ($P \leq 0.05$, Fisher's exact test, Additional File 3 Table S6). Of which two transmembrane orthogroups, that contained genes with these domains, also encoded for the laccase CAZy enzymes AA1_1, AA1_2, and AA1_3 (OG0005604, $\omega = 0.38$; OG0002895, $\omega = 0.22$) (Additional File 2 Table S1).

There was limited positive selection among predicted core effector genes (Fig. 4B). Only two predicted effector genes (Fig. 4A), corresponding to a proportion of 2.4% of the 84 predicted

effector genes examined (Table 2, Fig. 4B), had evidence of positive selection. Suggesting that core effectors might not be under pressure to evolve to overcome host defenses. These two predicted effector genes (OG0003219, $\omega = 0.76$, EffectorP mean score = 0.90 ± 0.028; OG0006565, $\omega = 1.96$, EffectorP mean score = 0.78 ± 0.051) did not have any associated protein domains (Additional File 2 Table S1, Additional File 3 Table S4). We also did not observe any evidence of positive selection in the 10 known virulence factors of *C. purpurea* (Mey *et al.* 2001, 2002; Oeser *et al.* 2002; Scheffer *et al.* 2005a, 2005b; Giesbert *et al.* 2008; Rolke *et al.* 2008; Bormann and Tudzynksi 2009) (Additional File 2 Table S1, Additional File 3 Table S4). In addition, we found no domain enrichment in positively selected secreted genes and CAZys. Peptidase (MEROP) genes only showed enrichment in an alpha/beta hydrolase fold domain (Additional File 3 Table S6).

Overall, our results revealed a lack of significant positive selection on predicted core effector genes, but a larger proportion of core unclassified and secondary metabolite genes with signatures of positive selection (Fig. 4). It should be noted that secondary metabolite genes also showed the highest number of codons per gene with signatures of positive selection, as determined by the Bayes Empirical Bayes (BEB) algorithm integrated into PAML, however, we did not observe significant differences between gene classifications (Fig. 4C).

Recombination landscape:

Recombination is also an important potential driver of genome evolution and plays a central role in the adaptability of parasitic organisms to overcome host defenses (Morran *et al.* 2011). Our genome-alignments contained 154 of the original 191 scaffolds (Table 3). The 37 missing scaffolds totaled 222,918 bp (average lengths = 6,192 ± 5,676 bp) and corresponded to

246 59 genes. Thirty-one of the missing scaffolds contained genes that were only part of the
247 accessory genome of which six scaffolds contained two or more genes (Additional File 3 Table
248 S8), suggesting that these scaffolds represent blocks of genetic material that could be lost or
249 gained from isolate to isolate. Most of the genes found on these scaffolds encoded conversed
250 domains associated with either reverse transcriptase, integrases, or helicases (Additional File 3
251 Table S8), which suggest unplaced repetitive content. Although, one scaffold (scaffold 185) did
252 possess a gene encoding a conserved domain for a centromere binding protein (Additional File 3
253 Table S8). Together these observations could indicate the potential for dispensable
254 chromosomes, as dispensable and mini-chromosomes often contain higher repetitive content
255 (Peng *et al.* 2019).

256 From our shared alignments we recovered 1,076,901 biallelic SNPs corresponding to a
257 median nucleotide diversity (Watterson's θ) of 0.01196 and a Tajima's D of -0.82522 calculated
258 from 10 kb non-overlapping windows (Table 3). The resulting SNPs were used to infer the
259 population recombination rate (ρ) from the linkage disequilibrium between SNPs based on *a*
260 *priori* specified population mutation rate θ, which was set to 0.01 based on our nucleotide
261 diversity (Watterson's θ) (Table 3) (Stukenbrock and Dutheil 2018a). The *C. purpurea* genome
262 recombination landscape was highly variable as some scaffolds showed highly heterogenous
263 landscapes, other scaffolds showed intermixed large peaks of recombination, while others still
264 had more constantly sized peaks across the regions (Fig. 5, Additional File 1 Fig. S6). Overall,
265 the mean genomic population recombination rate in *C. purpurea* was $\rho = 0.044$. We also
266 examined recombination in specific sequence features and gene type through comparison of
267 mean population recombination rates in exons, introns, 500-bp upstream and downstream of the
268 coding DNA sequence, and intergenic regions based on the annotation of the reference genome

(strain 20.1). The distribution of population recombination rates were comparable across different gene features and gene functional categories, although, some significant differences were observed (Fig. 6). In general, we found upstream regions to have the lowest recombination rates, while downstream regions have the highest recombination rates (Fig. 6). The decreased recombination in upstream regions might be the result of mechanisms trying to conserve promotor regions. This trend was observed across different functional gene categories, except in predicted effector genes where exons showed the highest recombination rates and downstream regions with the lowest, although these were not significantly different (Fig. 6B). Across functional categories, secreted genes and transmembrane genes showed the highest recombination rates within each gene feature but were not always significantly different (Fig. 6C).

Due to the observation of paralogs (Fig. 1) and evidence of tandem gene duplication in *C. purpurea* (Wyka *et al.* 2020a) we investigated the extent recombination might have influenced these events. We found that duplicated genes had lower population recombination rates than all other genes within the genome (Fig. 6D), suggesting that other factors are influencing gene duplication. Due to the absence of RIP (Wyka *et al.* 2020a), transposable elements (TEs) are likely a contributing factor. To investigate the association of duplicated genes with TEs we calculated the average distance of genes to long terminal repeat (LTR) retrotransposons and the average number of flanking LTRs. Results showed duplicated genes were significantly closer to LTRs and had significantly more flanking LTRs than predicted effector and other genes ($P <$ 0.0001, multi-test corrected Mann-Whitney U Test, Additional File 1 Fig. S7).

As we observed distinct peaks of recombination (Fig. 5, Additional File 1 Fig. S6), we further utilized LDhot to call statistically significant recombination hotspots by analysis of the

292 intensity of recombination rates in 3 kb (1 kb increments) windows compared to background
293 recombination rates in 20 kb windows (Auton *et al.* 2014; Wall and Stevison 2016; Stukenbrock
294 and Dutheil 2018a). After implementing a cuttoff of $\rho \geq 5$ and length of 20 kb (Wall and
295 Stevison 2016) we retained only five recombination hotspots, ranging from 11 kb to 18.5 kb in
296 length (Fig. 7). We observed a recombination hotspot located between the *lpsA1* and *lpsA2* genes
297 of the ergoline biosynthetic cluster, suggesting that this gene duplication event was likely the
298 result of recombination (Fig. 7D). Association of gene functional category and TEs within
299 hotspots varied between region. Some hotspots showed a greater association with duplicated
300 genes and TEs (Fig. 7B-D), while others showed a lower association (Fig. 7A and 7E). In
301 general, genes with conserved protein domains showed the highest presence within hotspots
302 (Additional File 1 Fig. S8). It should be noted that some unclassified genes and genes with
303 conserved protein domains associated with hotspots were also found to be overlapping regions
304 identified as repeats (Fig. 7A-C and 7E). Protein domains found within these genes were
305 associated with ankyrin (IPR002110) and tetratricopeptide (IPR013026) repeats. Only 5 of the
306 846 duplicated genes (reported in Wyka *et al.* 2020a) found throughout the reference genome
307 were located within predicted recombination hotspots (Fig. 7, Additional File 1 Fig. S8). While
308 Wyka *et al.* (2020a) showed that gene cluster expansion was prevalent among predicted
309 effectors, we only found one non-duplicated predicted effector (CCE30212.1) located within a
310 recombination hotspot (Fig. 7C). Together these results suggest that while recombination may
311 result in important gene duplication, it is not the primary driver of gene duplication within *C.*
312 *purpurea*.
313
314 **Discussion:**

Our establishment of a *Claviceps purpurea* pangenome from 24 isolates, as well as, the detection of core genes with signatures of positive selection and analysis of the recombination landscape have provided knowledge into how high recombination rates, gene duplication, and selection of secondary metabolite genes are driving the genomic evolution and adaptation of the species.

The pangenome of *C. purpurea* reveals a large accessory genome with 37.78% accessory orthogroups (27.05% accessory + 10.73% singleton) in comparison to four model fungal pangenomes (*Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans*, and *Aspergillus fumigatus*), which found around 9 – 19% of their genes in the accessory genome (McCarthy and Fitzpatrick 2019). Our results are more comparable to the pangenome of the fungal pathogen *Zymoseptoria tritici* which had an accessory genome comprised of 40% (30% accessory + 10% singleton) of genes (Badet *et al.* 2020). Similar to *C. purpurea*, *Zymoseptoria tritici* is a globally distributed biotrophic fungal pathogen of grasses, notably wheat, suggesting that fungal species with similar life strategies, hosts, and ecological environments could possess comparable pangenome structures as they are under similar evolutionary pressures. Similar factors of lifestyle, effective population size, and habitat have been reported to influence pangenome sizes in bacteria (Mclnerney *et al.* 2017). In fact, *C. purpurea* and *Z. tritici* both experienced enrichment of predicted effector orthogroups in the accessory genome and enrichment of carbohydrate-active enzymes (CAZys) orthogroups in the core genome (Fig. 2) (Badet *et al.* 2020), conveying a comparable similarity between gene functions within pangenome structure regarding the pathogenic lifestyle of these organisms. In addition, Badet *et al.* (2020) suggested that the large accessory genome of *Z. tritici* is likely maintained due to TE activity and a large effective population size as a result of observations of high SNP density,

338 rapid decay in linkage disequilibrium, and high recombination rates (Croll *et al.* 2015; Hartmann
339 *et al.* 2017; Stukenbrock and Dutheil 2018a). The same mechanisms could also explain the large
340 accessory genome observed in *C. purpurea*.
341 We observed an abundance of orthogroups containing paralogs (8.6%), potentially due to
342 a lack of RIP (Wyka *et al*. 2020a). This presence of gene duplication and association with LTR
343 retrotransposons (Additional File 1 Fig. S7) could be contributing to the large size of the
344 accessory genome, potentially through pseudogenization and/or neofunctionalization. In fact,
345 unclassified genes had the highest ω (dN/dS) ratios (Fig. 3) and the highest proportion of genes
346 with signatures of positive selection (Fig. 4). While this analysis was only conducted on single-
347 copy core genes, it suggests that some of the unclassified accessory genes (Fig. 2H) are
348 undergoing similar evolutionary trends. In addition, the abundance of duplication in accessory
349 unclassified genes (Wyka *et al.* 2020a) and their small sizes (Additional File 1 Fig. S2) can
350 further suggest the presence of pseudogenization and/or neofunctionalization. Badet *et al.* (2020)
351 suggested that TEs were likely contributing to *Z. tritici* accessory genome due to their
352 correlations of TE content with genome size and observations of transcribed TEs. We observed a
353 similar correlation of TE content with genome size ($P = 0.004$, Adj. $R^2 = 0.28$), however, our
354 genome sizes and TE content (30.5 Mb – 32.1 Mb, 8.42% - 10.87%, respectively) were not as
355 variable as in *Z. tritici*, which also had a twofold higher TE content (Badet *et al.* 2020). This
356 suggests that TEs play a more important role in *Z. tritici* genome expansion, however, only 0.2%
357 of the orthogroups in *Z. tritici* contained paralogs suggesting that gene duplication is not as
358 common in *Z. tritici* as it is in *C. purpurea* (8.6% paralogs). The lack of gene duplication in *Z.*
359 *tritici* is likely due to the presence of RIP (Testa *et al.* 2015), which should also reduce TE
360 expansion through silencing (Galagan *et al.* 2003, 2004; Urquhart *et al.* 2018). While we lack

RNAseq data to observe TE transcription within *C. purpurea*, observations of TEs with 0% divergence in *C. purpurea* (Wyka *et al.* 2020a) suggest recent TE activity. The observed lack of recombination associated with duplicated genes (Fig. 6D) and association of duplicated genes with LTR transposons (Additional File 1 Fig. S7) would suggest that gene duplication in *C. purpurea* is mediated in part by transposon activity.

Furthermore, we identified 37 missing scaffolds in our population genome alignment with 31 of these containing genes only present in the accessory genome, suggesting the potential for blocks of DNA that could be lost/gained between isolates. Of these accessory scaffolds 15 contained genes encoding conversed domains associated with either reverse transcriptase, integrases, or helicases and one scaffold possessed a gene encoding a conserved domain for a centromere binding protein (Additional File 3 Table S8). Together these could indicate the potential for dispensable mini-chromosomes, as dispensable and mini-chromosomes often contain higher repetitive content (Peng *et al.* 2019). However, even the combination of all 37 missing scaffolds (0.22 Mb) would represent the smallest mini-chromosome known in plant pathogens; 3-fold smaller than *Leptosphaeria maculans* (Balesdent *et al.* 2013), 2-fold smaller than *Nectria haematococca* (Mahmoud and Taga 2012), and 7-fold smaller than *Magnaporthe oryzae* (Peng *et al.* 2019). Many of these scaffolds contained repeated N's sequences from scaffolding (Schardl *et al.* 2013) and increased repeat content (Additional File 3 Table S8) suggesting that our Illumina based genomes might not have captured the true nature of these scaffolds. Therefore, we did not process these elements further but believe that these are an important aspects of *C. purpurea* evolution and should be a focal point of future research with the advantage of long-read sequencing to more confidently understand their function. Due to these transcriptase rich unplaced scaffolds, the lack of RIP, association of duplicated genes with

384 transposons, and observation of TEs with 0% divergence (Wyka *et al.* 2020a), we believe
385 transposons and/or transcriptases are influencing gene duplication in *C. purpurea*.
386 Due to the potential for transposon mediated gene duplication, it was remarkable to find
387 relatively low TE content (~8 - 10%) within *C. purpurea*, especially in the absence of RIP. Other
388 genomic mechanism, such as recombination, may help to limit TE expansion and increases in
389 genome size. Tiley and Burleigh (2015) found a strong negative correlation between global
390 recombination rate, genome size and LTR retrotransposon proportion across 29 plant species,
391 indicating that higher recombination rates actively reduce genome size likely through the
392 removal of LTR elements. A similar function may be affecting LTR content in *C. purpurea*,
393 which would explain the observed differences in LTR content between *Clavicpes* section
394 *Claviceps* (low LTR content, RIP absent) and *Claviceps* sections *Pusillae*, *Paspalorum*, and
395 *Citrinae* (high LTR content, RIP present) (Wyka *et al.* 2020a).
396 On average we observed a twofold higher mean population recombination rate (ρ =
397 0.044) in *C. purpurea* than *Z. tritici* (ρ = 0.0217) and tenfold higher than *Z. ardabiliae* (ρ =
398 0.0045) (Stukenbrock and Dutheil 2018a). As ρ is a function of effective population size and
399 recombination rate per site ($\rho = 2N_e \times r$), these increases could be the result of the increment in
400 recombination rate per site (r) and/or effective population size (N_e). Differences in ρ between the
401 two *Zymoseptoria* species was postulated to be due to increased recombination rates per site as it
402 was found that the nucleotide diversity (Watterson's $\theta = 2\ N_e \times \mu$, where μ is mutation rate) was
403 1.6 times higher in *Z. tritici* (0.0139) than *Z. ardabiliae* (0.00866). Under an assumption that
404 both *Z. tritici* and *Z. ardabiliae* have comparable mutation rates, N_e of *Z. tritici* would only be 1.6
405 times higher than *Z. ardabiliae*, therefore, the 5 fold higher ρ would likely be caused by higher
406 recombination rates per site (Stukenbrock and Dutheil 2018a). Our observed Watterson's θ of

0.012 in *C. purpurea* (Table 2) is comparable to *Z. tritici*, suggesting that if mutation rates and effective populations sizes are comparable than the twofold increase in ρ is likely influenced by higher recombination rates per site in *C. purpurea*. Although, *Z. tritici* is a heterothallic organism while *C. purpurea* is homothallic (Esser and Tudzynski 1978) but *C. purpurea* also frequently out-crosses in nature (Amici *et al.* 1967; Tudzynski 2006), suggesting that these factors may provide a difference in effective population sizes between these organisms. In addition, mutation rates might differ between *C. purpurea* and *Z. tritici* for several reasons. Selection pressure associated with agriculture control methods could be driving the mutation of *Z. tritici*, which is subjected to multiple annual fungicide treatments (Torriani *et al.* 2015) and multiple cultivars with various qualitative and quantitative resistance sources (Brown *et al.* 2015). In contrast, control of *C. purpurea* is focused on cultural practices as fungicides have proven inefficient and no resistance crop germplasm has been identified (Menzies and Turkington 2015). While fungicides and crop resistance affect the population structure of *Z. tritici* (Estep *et al.* 2015; Hayes *et al.* 2016; Welch *et al.* 2018), it is plausible to believe they might affect mutation rate or select for strains with a higher mutation or recombination rates. However, we are unaware of any study that has directly examined whether fungicides or crop resistance can have direct or indirect effects on mutation rates. An alternative, and more plausible, hypothesis to explain an increased mutation rate in *Z. tritici* would be associated with the function of RIP, which identifying repeat/duplicated sequences within a genome and introduces C:G to T:A mutations to effectively silence these regions (Galagan *et al.* 2003, 2004; Urquhart *et al.* 2018). It has also been reported that RIP can "leak" into neighboring non-repetitive regions and introduce mutations, thus, accelerating the rate of mutations, particularly those in closer proximity to repeat regions (Fudal *et al.*, 2009; Hane *et al.*, 2015; Van de Wouw *et al.*, 2010). If the mutation rate is increased in *Z.*

tritici, either due to RIP "leakage" or selective pressure from fungicides or host resistance the nucleotide diversity in *Z. tritici* could be the result of high mutation rates, whereas the nucleotide diversity in *C. purpurea* could be influenced by higher effective population size and/or recombination rates per site. Higher recombination rates were found to increase the efficacy of purifying selection in both plants (Tiley and Burleigh 2015) and *Z. tritici* (Grandaubert *et al.* 2019). Similarly, *C. purpurea* had an overall trend of purifying selection with skewness towards lower ω values (Fig. 3) and an observed correlation of higher population recombination rates around genes with lower ω ratios (Additional File 1 Fig S9), further suggesting the potential for higher recombination rates in *C. purpurea*.

Additional support, for higher recombination rates per site in *C. purpurea*, could be extrapolated from recombination hotspots, or lack thereof. While we observed evidence of a heterogenous recombination landscapes with several scaffolds showing large peaks in population recombination rates (Fig. 5, Additional File 1 Fig. S6), we only predicted five recombination hotspots (Fig. 7), which is in stark contrast to the ~1,200 hotspots identified in *Z. tritici* (Stukenbrock and Dutheil 2018b, *Updated dataset*). On average, we did observe higher population recombination rates across scaffolds compared to the rates observed across chromosomes of *Zymoseptoria* (Stukenbrock and Dutheil 2018a), suggesting that the background recombination rate in *C. purpurea* is higher and "flatter", potentially limiting the detection of hotspots (Auton *et al.* 2014). Overall, this indicates that *C. purpurea* exhibits high recombination rates per site, which potentially helps defend against TE expansion.

While these higher recombination rates are likely influencing the trend of strong purifying selection observed in *C. purpurea*, it might not be the sole factor responsible for the low number of predicted core effector genes with signatures of positive selection (Fig. 4). Wäli

453 *et al.* (2013) classified *C. purpurea* as a conditional defense mutualist with its plant host, as they
454 found that sheep avoided grazing infected grasses and observed that infection rates were higher
455 in grazed pastures compared to ungrazed fields. Other researchers have observed neutral to
456 positive effects of seed set, seed weight, and plant growth on infected plants compared to
457 uninfected plants (Raybould *et al*. 1998; Fisher *et al.* 2007; Wäli *et al.* 2013; Wyka *et al.* 2020
458 *Unpublished PhD Dissertation*). These factors, along with the broad host range of *C. purpurea*
459 (400+ grass species) and lack of known crop resistance (R) genes, could suggest a lack of strong
460 selection for resistance to *C. purpurea* in grass species (Wäli *et al.* 2013). This could help
461 explain the lack of positive selection observed in predicted core effector genes, implying that
462 effectors are not under strong selection pressure to compete in the evolutionary arms race against
463 host defense. However, it should be noted that positive selection analyses are computed from
464 single-copy core orthologs. Observations of significant enrichment of predicted effector genes in
465 the accessory genome of *C. purpurea* and duplication of effector gene cluster (Wyka *et al.*
466 2020a) could implicate their role in diversity of infection potential (Sánchez-Vallet *et al.* 2018),
467 however, no host specific races of *C. purpurea* have been identified.

468 *Claviceps purpurea*, which is suggested to have an ancestral state of plant endophytism
469 (Píchová *et al.* 2018) is also closely related to several mutualistic grass endophytes (i.e. *Epichloe*,
470 *Balansia*, *Atkinsonella*) which have been known to provide beneficial aspects to their hosts
471 mostly through production of secondary metabolites and plant hormones (Clay 1988; Song *et al.*
472 2016; Xia *et al.* 2018). *Claviceps purpurea* is well-known for its secondary metabolite
473 production and, as we observed, had the second highest proportion of genes with positive
474 selection signatures and the highest number of codons under selection per gene (Fig. 4B and 4C).
475 We also observed two orthogroups with signatures of positive selection containing domains for

476 laccase CAZy enzymes - with some laccases facilitating the biosynthesis of melanin in fungi
477 (Lee *et al.* 2019) - and selection signatures on the cytochrome P450 associated with fungal
478 cytokinin biosynthesis (Hinsch *et al.* 2015). Secondary metabolites are known to increase stress
479 tolerance in fungi (i.e. against UV radiation, oxidative stresses, or colder climates) as has been
480 shown with several groups of pigments, such as melanins and carotenoids (Avalos and Carmen
481 Limon 2015). Therefore, the evolution of secondary metabolites in *C. purpurea* (i.e. ergot
482 alkaloids, ergochromes, or other pigments) can theoretically increase fitness through altering
483 infection potential, stress tolerance, or antimicrobial resistance (Píchová *et al.* 2018; Pusztahelyi
484 *et al.* 2019). The difference in the proportion of secondary metabolites genes under positive
485 selection pressure (such as the ergoline biosynthesis cluster), compared to predicted effectors,
486 indicates that the evolution of secondary metabolite genes in *C. purpurea* is more important to
487 the success of the species than adaptation of core effector proteins. This is in contrast to many
488 fungal plant pathogens of cereal crops, such as *Z. tritici* and the rust fungi in the genus *Puccinia*,
489 that rely on adaptation and diversification of effector proteins for success, particularly due to
490 breeding of crop varieties with R genes (Sánchez-Vallet *et al.* 2018; Badet *et al.* 2020). The
491 selective pressure on secondary metabolites in *C. purpurea* could help explain its evolutionary
492 history as it was recently postulated that evolution of *Claviceps* section *Claviceps*, of which *C.*
493 *purpurea* resides, occurred tandemly with the radiation of the core Pooideae (Poeae, Triticeae,
494 Bromeae, and Littledaleeae) and was associated with adaptation and diversification to cooler,
495 more open habitats (Kellogg 2001; Sandve and Fjellheim 2010; Píchová *et al.* 2018; Wyka *et al.*
496 2020a). In addition, the speciation among *C. purpurea* and closely related species demonstrate
497 varied levels of adaptation to ecological niches (Pažoutová *et al.* 2000, 2002, 2015; Douhan *et al.*
498 2008; Negård *et al.* 2015; Shoukouhi *et al.* 2019; Liu *et al. Submitted*). Similar evolutionary

499 trends towards positive selection of secondary metabolites could be influencing the divergence of
500 these species as well. In fact, all members of *Claviceps* section *Claviceps* had genomes that lack
501 RIP, exhibit gene duplication, and have comparable TE content (Wyka *et al.* 2020a), suggesting
502 that the genomic mechanisms identified in this study might be characteristic of section *Claviceps*
503 as a whole.
504
505 **Conclusion:**
506 Overall, we observed that the *Claviceps purpurea* pangenome is composed of a large
507 accessory genome that is likely influenced by a large effective population size, high
508 recombination rates, and TE mediated gene duplication. Pseudogenization and
509 neofunctionalization might also be contributing due to the observed TE activity, observations of
510 higher ω ratios, signatures of positive selection in core single-copy unclassified genes, and small
511 size of many accessory unclassified genes. Due to a lack of RIP, prolific TE expansion is likely
512 controlled by high recombination rates, which subsequently may be influencing the overall trend
513 of purifying selection. However, secondary metabolites genes were found to have the highest
514 rates of positive selection on codons within genes, indicating that these genes are a primary
515 factor affecting the diversification of the species into new ecological niches and to potentially
516 help maintain its global distribution and broad host range.
517
518 **Materials and Methods:**
519
520 *Genome data:*

Haploid genome data from a collection of 24 isolates was utilized in this study to provide a comprehensive analysis of *Claviceps purpurea*. The 32.1 Mb reference genomes of *C. purpurea* strain 20.1 was sequenced in 2013 using a combination of single and paired-end pyrosequencing (3 kb fragments) resulting in a final assembly of 191 scaffolds (Schardl *et al.* 2013; NCBI: SAMEA2272775). The remaining 23 isolates were recently sequenced, assembled, and annotated in Wyka *et al.* (2020a: NCBI BioProject: PRJNA528707), representing a collection of isolates from USA, Canada, Europe, and New Zealand (Table 1). The reference genome was subject to an amino acid cutoff of 50 aa to match the other 23 isolates. In this study, we report the pangenome of *C. purpurea*, analysis of the population genomic recombination, and the landscape of genes under positive selection.

Gene functional and transposable element (TE) annotations utilized were those reported in Wyka *et al.* (2020a) and datasets Wyka *et al.* (2020b *Dryad dataset*). In brief, secondary metabolite clusters were predicted using antiSMASH v5 (Blin *et al.* 2019), with all genes belonging to identified clusters classified as "secondary (2°) metabolites". Functional domain annotations were conducted using InterProScan v5 (Jones *et al.* 2014), HMMer v3.2.1 (Wheeler and Eddy 2013) search against the Pfam-A v32.0 and dbCAN v8.0 CAZYmes databases, and a BLASTp 2.9.0+ search against the MEROPs protease database v12.0 (Rawlings *et al.* 2018). Proteins were classified as secreted proteins if they had signal peptides detected by both Phobius v1.01 (Käll 2007) and SignalP v4.1 (Nielsen 2017) and did not possess a transmembrane domain as predicted by Phobius and TMHMM v2.0 (Krogh *et al.* 2001). Effector proteins were identified by using EffectorP v2.0 (Sperschneider *et al.* 2018) on the set of secreted proteins for each genome. Transmembrane proteins were identified if both Phobius and TMHMM detected transmembrane domains. Transposable elements fragments were identified following procedures

for establishment of *de novo* comprehensive repeat libraries set forth in Berriman *et al.* (2018) through a combined use of RepeaModeler v1.0.8 (Smit & Hubley 2015), TransposonPSI (Hass 2010), LTR_finder v1.07 (Xu & Wang 2007), LTR_harvest v1.5.10 (Ellinghaus *et al.* 2008), LTR_digest v1.5.10 (Steinbiss *et al.* 2009), Usearch v11.0.667 (Edgar 2010), and RepeatClassifier v1.0.8 (Smit & Hubley 2015) with the addition of all curated fungal TEs from RepBase (Bao *et al.* 2015). RepeatMasker v4.0.7 (Smit *et al.* 2015) was then used to identify TE regions and soft mask the genomes. These steps were automated through construction of a custom script, TransposableELMT (https://github.com/PlantDr430/TransposableELMT) (Wyka *et al.* 2020a, 2020b).

Pangenome analysis:

The pangenome was constructed using OrthoFinder v2.3.3 (Emms *et al.* 2019), on all genes identified from the 24 genomes, to infer groups of orthologous gene clusters (orthogroups). OrthoFinder was run using BLASTp on default settings. For downstream analysis, gene clusters were classified as secreted, predicted effectors, transmembrane, secondary (2°) metabolites, carbohydrate-degrading enzymes (CAZys), proteases (MEROPs), and conserved domain (conserved) clusters if ≥ 50% of the strains present in a gene cluster had at least one protein classified as such. Gene clusters not grouped into any of the above categories were categorized as unclassified.

Core and pangenome size curves were extrapolated from resampling of 24 random possible combinations for each pangenome size of 1 - 24 genomes and modelled by fitting the power law regression formula: $y = Ax^B + C$ using the curve_fit function in the Python module

566 Scipy v1.4.1. These processes were automated through the creation of a custom python script
567 (https://github.com/PlantDr430/FunFinder_Pangenome).
568
569 *Positive selection*:
570 To investigate the positive selection landscape of genes we collected a total of 6,243
571 single-copy orthologs across all 24 genomes (See Table 2 for detailed report). For each ortholog
572 cluster sequences were aligned using MUSCLE v3.8.1551 (Edgar 2004) on default settings and
573 values of dN, dS, and dN/dS (omega, ω) were estimated using the YN00 (Yang and Nielsen
574 2000) method in PAML v4.8 using default parameters. Each ortholog was then individually
575 examined for evidence of positive selection. Guide trees were generated for each ortholog cluster
576 using FastTree version 2.1.10 SSE3 and positive selection was detected using the CodeML
577 algorithm (Yang 2007a) in PAML v4.8 with parameters: NSites = 0 1 2 3 7 8, CodonFreq = 2,
578 seqtype = 1, kappa = 0.3, omega = 1.3, ncatG = 10. Due to high average nucleotide similarities in
579 pairwise BLASTn searches within each ortholog (Additional File 1 Fig. S4) we utilized a
580 stringent filtering method to enhance our confidence in the selection of genes with positive
581 selection signatures. Orthologs were only identified as being under positive selection if they were
582 significant at $\alpha \leq 0.01$ using a likelihood ratio test (df -2, $\chi 2$ critical value = 9.13) in both the M7
583 vs. M8 and M2 vs. M1 model comparisons. In addition, orthologs also needed to contain at least
584 one specific amino acid residue significantly ($\alpha \leq 0.01$) identified as being under positive
585 selection using the Bayes Empirical Bayes algorithm integrated into PAML (Yang 2007a), in
586 both the M8 and M2 models.
587 For statistical purposes, each gene cluster was only characterized by one functional
588 category in the order displayed in Table 2 (i.e. secreted genes are those not already classified as

589 effectors, etc). After filtering for positive selection, gene functional categories were examined for
590 enrichment of Pfam, Iprscan, MEROPs, CAZy, and smCOGs domains, as well as, gene ontology
591 (GO) terms (See Methods section *Statistical analyses and plotting*).
592
593 *Genome alignment, SNP calling, and recombination*:
594 Procedures followed Stukenbrock and Dutheil (2018a), for creation of a fine-scale
595 recombination map of fungal organisms and identification of recombination hotspots. A brief
596 description will be provided below, for a more detailed methodology and explanation of
597 algorithms refer to Stukenbrock and Dutheil (2018a), Auton *et al.* (2014), and Wall and Stevison
598 (2016).
599 LastZ and MultiZ from the TBA package (Blanchette *et al.* 2004) was used to create the
600 population genome alignment projected against the reference genome, *C. purpurea* strain 20.1
601 (Schardl *et al.* 2013). Alignments in MAF format were filtered using MafFilter v.1.3.1 (Dutheil
602 *et al.* 2014) following Stukenbrock and Dutheil (2018a). Final alignments were merged
603 according to the reference genome and subsequently divided into nonoverlapping windows of
604 100 kb. MafFilter was additionally used to compute genome-wide estimates of nucleotide
605 diversity (Watterson's θ) and Tajima's D in 10 kb windows. Single nucleotide polymorphisms
606 (SNPs) were called by MafFilter from the final alignment. Principal Component Analysis (PCA)
607 and a Maximum-Likelihood phylogeny were conducted with fully resolved biallelic SNPs (Table
608 3) using the R package SNPRelate v1.18.1 (Zheng *et al.* 2012) and RAxML v8.2.12 (Stamatakis
609 2014) using GTRGAMA and 1000 bootstrap replicates, respectively.
610 The following process was automated through the creation of a custom python script
611 (https://github.com/PlantDr430/CSU_scripts/blob/master/Fungal_recombination.py). LDhat

(Auton and McVean 2007) was used to estimate population recombination rates (ρ) from the filtered alignment using only fully resolved biallelic positions. A likelihood table was created for the θ value 0.01, corresponding to the genome-wide Watterson's θ of *C. purpurea* (Table 3; Julien Dutheil *per comm*), and LDhat was run with 10,000,000 iterations, sampled every 5000 iterations, with a burn-in of 100,000. The parameter ρ relates to the actual recombination rate in haploid organism through the equation $\rho = 2N_e \times r$, where N_e is the effective population size and r is the per site rate of recombination. However, without knowledge of N_e we cannot confidently infer r and thus sought to avoid the bias of incorrect assumptions. Therefore, we reported the population recombination rate (ρ).

Resulting recombination maps were filtered to remove pairs of SNPs for which the confidence interval of the recombination estimate was higher than two times the mean (Stukenbrock and Dutheil 2018a). Average recombination rates were calculated, in regions, by weighing the average recombination estimate between every pair of SNPs by the physical distance between the SNPs. Using the reference annotation file (Schardl *et al.* 2013), we calculated the average recombination rates for features in each gene: 1) exons, 2) introns, 3) 500 bp upstream, and 4) 500 bp downstream with a minimum of three filtered SNPs. Flanking upstream and downstream regions correspond to the 5′ and 3′ regions for forward stranded genes and the 3′ and 5′ regions for reverse stranded genes. We also calculated the average recombination rate for each intergenic region between the upstream and downstream regions of each gene. Introns were added to the GFF3 file using the GenomeTools package (Gremme *et al.* 2013). The original recombination maps produced from LDhat (Julien Dutheil *per comm*) were converted from bp to kb format for use in LDhot (Auton *et al.* 2014) to detect recombination hotspots 1000 simulations and --windlist 10 to create 20 kb background windows (Wall and

635 Stevison 2016). Only hotspots with a value of ρ between 5 and 100 and width < 20 kb were
636 selected for further analysis (Auton *et al.* 2014; Wall and Stevison 2016; Stukenbrock and
637 Dutheil 2018a).
638
639 *Statistical and enrichment analyses:*
640 Statistics and figures were generated using Python3 modules SciPy v1.3.1, statsmodel
641 v0.11.0, Matplotlib v3.1.1, and seaborn v0.10.0. All multi-test corrections were performed with
642 Benjamini-Hochberg false discovery rate procedure. Enrichment analyses were tested using
643 Fischer's Exact test with a cutoff $\alpha = 0.05$. Uncorrected p-values were corrected using
644 Benjamini-Hochberg and Bonferroni multi-test correction with a false discovery rate (FDR)
645 cutoff of $\alpha = 0.05$. Corresponding p-values from correction tests were averaged together to get a
646 final p-value. Enrichment was performed on protein domain names and GO terms. Orthogroups
647 were only associated with a domain or GO term if ≥ 50% of the strains present in the gene
648 cluster had one gene with the term. This process was automated through creation of a custom
649 python script (https://github.com/PlantDr430/CSU_scripts/blob/master/Domain_enrichment.py).
650
651 **Acknowledgements:**
652 We would like to thank Julien Dutheil for his assistance in understanding the procedures
653 for estimating fungal recombination rates using LDhat and LDhot.
654
655 **Funding:**
656 This work is supported by the Agriculture and Food Research Initiative (AFRI) National
657 Institute of Food and Agriculture (NIFA) Fellowships Grant Program: Predoctoral Fellowships

658 grant no. 2019-67011-29502/project accession no. 1019134 from the United States Department
659 of Agriculture (USDA), and by the American Malting Barley Association grant no. 17037621.
660 Dr. Broders is supported by the Simon's Foundation Grant number 429440 to the Smithsonian
661 Tropical Research Institute.

662

663 **Data availability:**

664 Most of the relevant data are within the manuscript and supporting files. Additional raw
665 datasets and scripts are available on Dryad: Wyka, Stephen *et al.* (2020), A large accessory
666 genome, high recombination rates, and selection of secondary metabolite genes help maintain
667 global distribution and broad host range of the fungal plant pathogen *Claviceps purpurea*, v1,
668 Dryad, Dataset, doi: 10.5061/dryad.6hdr7sqxp.

669

670 **Author Contributions:**

671 The project was conceived and designed by S.A.W., S.J.M., and K.B.; S.A.W. performed
672 the research, bioinformatic and software workflows, and analyzed and visualized the data with
673 technical troubleshooting from S.J.M.; M.L., V.N., and K.B. provided management, supervision,
674 research advice, and editorial contributions; S.A.W. wrote the paper with contributions from all
675 other authors.

676

677 **Competing interests:**

678 The authors have declared that no competing interests exist.

679

680 **References:**

[Dryad dataset]* Wyka SA, Mondo SJ, Liu M, Dettman J, Nalam V, Broders KD (2020b) Whole genome comparisons of ergot fungi reveals the divergence and evolution of species within the genus *Claviceps* are the result of varying mechanisms driving genome evolution and host range expansion. v4, Dryad, Dataset, doi: 10.5061/dryad.18931zcsk.

[Gitlab dataset]* Stukenbrock EH, Dutheil JY (2018b) Data and scripts for: Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. Available at: https://gitlab.gwdg.de/molsysevol/ZtPopRec.

Araki H, Tian D, Goss EM, Jakob K, Halldorsdottir SS, Kreitman M, *et al.* (2006) Presence/absence polymorphism for alternative pathogenicity islands in *Pseudomonas viridiflava*, a pathogen of Arabidopsis. *Pnas* **103,** 5887-92.

Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. *Genome Research* **17,** 1219-1227.

Auton A, Myers S, and McVean G (2014) Identifying recombination hotspots using population genetic data. a*rXiv*: 1403.4264.

Avalos J, Carmen Limon M (2015) Biological roles of fungal carotenoids. *Current Genetics* **61,** 309–324

Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D (2020) A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. *BMC Biology* **18,** 12.

Balesdent M-H, Fudal I, Ollivier B, *et al.* (2013) The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytologist* **198,** 887 – 98.

704 Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in
705 eukaryotic genomes. *Mobile DNA* **6,** 11.
706 Berriman M, Coghlan A, Tsai IJ (2018) Creation of a comprehensive repeat library for newly
707 sequenced parasitic worm genome. *Protocol Exchange* doi: 101038/protex2018054.
708 Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA *et al.* (2004) Aligning multiple
709 genomic sequences with the threaded blockset aligner. *Genome Research* **14,** 708–715.
710 Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T (2019)
711 antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids*
712 *Research* **47,** W81-87.
713 Bormann J, Tudzynski P (2009) Deletion of Mid1, a putative stretch-activated calcium channel
714 in *Claviceps purpurea*, affects vegetative growth, cell wall synthesis and virulence.
715 *Microbiology* **155,** 3922-3933.
716 Brown JKM, Chartrain L, Lasserre-Zuber P, Saintenac C (2015) Genetics of resistance to
717 *Zymoseptoria tritici* and applications to wheat breeding. *Fungal Genetics and Biology* **79,** 33-
718 41.
719 Brynildsrud O, Gulla S, Feil EJ, Nørstebø SF, Rhodes LD (2016) Identifying copy number
720 variation of the dominant virulence factors msa and p22 within genomes of the fish pathogen
721 *Renibacterium salmoninarum*. *Microbial Genomics* **2,** e000055.
722 Clay K (1988) Fungal endophytes of grasses: a defensive mutualism between plants and fungi.
723 *Ecology* **69,** 10-16.
724 Croll D, Lendenmann MH, Stewart E, BA MD (2015) The impact of recombination hotspots on
725 genome evolution of a fungal plant pathogen. *Genetics* **201,** 1213-28.

726 Dopstadt J, Neubauer L, Tudzynski P, Humpf H-U (2016) The epipolythiodiketopiperazine gene
727 cluster in *Claviceps purpurea*: dysfunctional cytochrome p450 enzyme prevents formation of
728 the previously unknown clapurines. *PLoS ONE* **11,** e0158945.
729 Douhan GW, Smith ME, Huyrn KL, Westbrook A, Beerli P, Fisher AJ (2008) Multigene
730 analysis suggests ecological speciation in the fungal pathogen *Claviceps purpurea*.
731 *Molecular Ecology* **17,** 2276-2286.
732 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high
733 throughput. *Nucleic Acids Research* **32**, 1792-1797.
734 Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*
735 **26,** 2460-2461.
736 Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, a efficient and flexible software for de
737 novo detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 18.
738 Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative
739 genomics. *Genome Biology* **20,** 238.
740 Esser K, Tudzynski P (1978) Genetics of the ergot fungus *Claviceps purpurea*. I. Proof of a
741 monoecious life-cycle and segregation patterns for mycelial morphology and alkaloid
742 production. *Theoretical Applied Genetics* **53,** 145-149.
743 Estep LK, Torriani SFF, Zala M, Anderson NP, Flowers MD, McDonald BA, Mundt CC,
744 Brunner PC (2015) Emergence and early evolution of fungicide resistance in North American
745 populations of *Zymoseptoria tritici*. *Plant Pathology* **64,** 961-971.
746 Fisher AJ, DiTomaso JM, Gordon TR, Aegerter BJ, Ayres DR (2007) Salt marsh *Claviceps*
747 *purpurea* in native and invaded Spartina marshes in Northern California. *Plant Disease* **91,**
748 380–386.

749 Flieger M, Stodůlková E, Wyka SA, *et al.* (2019) Ergochromes: heretofore neglected side of
750 ergot toxicity. *Toxins* **11,** 439.
751 Fudal I, Ross S, Brun H, *et al.* (2009) Repeat-induced point mutation (RIP) as an alternative
752 mechanism of evolution towards virulence in *Leptosphaeria maculans*. *Molecular Plant-*
753 *Microbe Interactions* **22,** 932-941.
754 Galagan JE Selker EU (2004) RIP: the evolutionary cost of genome defense. *TRENDS in*
755 *Genetics* **20,** 417-423.
756 Galagan JE, Calvo SE, Borkovich KA *et al*. (2003) The genome sequence of the filamentous
757 fungus *Neurospora crassa*. *Nature* **422,** 859-868.
758 Giesbert S, Schurg T, Scheele S, Tudzynski P (2008) The NADPH oxidase Cpnox1 is required
759 for full pathogenicity of the ergot fungus *Claviceps purpurea*. *Molecular Plant Pathology* **9,**
760 317-327.
761 Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding
762 DNA sequences. *Molecular Biology and Evolution* **11,** 725-736.
763 Gordon A, McCartney C, Knox RE, *et al.* (2020) Genetic and transcriptional dissection of
764 resistance to *Claviceps purpurea* in the durum wheat cultivar Greenshank. *Theoretical and*
765 *Applied Genetics* doi: 10.1007/s00122-020-03561-9.
766 Grandaubert J, Dutheil JY, Stukenbrock EH (2019) The genomic determinants of adaptive
767 evolution in a fungal pathogen. *Evolution Letters* **3,** 299-312.
768 Gremme G, Steinbiss S, Kurtz S (2013) GenomeTools: a comprehensive software library for
769 efficient processing of structured genome annotations. *IEEE/ACM Transactions on*
770 *Computational Biology and Bioinformatics* **10,** 645–656.

771 Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP (2015) Repeat-induced point
772 mutation: a fungal-specific, endogenous mutagenesis process. In: van den Berg MA
773 Maruthachalam K, editors. *Genetic transformation systems in fungi*. Vol. 2. Springer
774 International Publishing, p. 55-68.
775 Hartmann FE, Rodríguez de la Vega RC, Brandenburg J-T, Carpentier F, Giraud T (2018) Gene
776 presence–absence polymorphism in castrating Anther-Smut fungi: recent gene gains and
777 phylogeographic structure. *Genome Biology and Evolution* **10,** 1298-314.
778 Hartmann FE, Sánchez-Vallet A, McDonald BA, Croll D (2017) A fungal wheat pathogen
779 evolved host specialization by extensive chromosomal rearrangements. *ISME Journal* **11,**
780 1189-204.
781 Hass B (2010) TransposonPSI. Available online at: http://transposonpsisourceforgenet.
782 Hayes LE, Sackett KE, Anderson NP, Flowers MD, Mundt CC (2016) Evidence of selection for
783 fungicide resistance in *Zymoseptoria tritici* populations on wheat in western Oregon. *Plant*
784 *disease* **100,** 483-489.
785 Hinsch J, Galuszka P, Tudzynski P (2016) Functional characterization of the first filamentous
786 fungal tRNA-isopentenyltransferase and its role in the virulence of *Claviceps purpurea*. *New*
787 *Phytologist* **211,** 980-992.
788 Hinsch J, Vrabka J, Oeser B, Novák O, Galuszka P, Tudzynski P (2015) *De novo* biosynthesis of
789 cytokinins in the biotrophic fungus *Claviceps purpurea*. *Environmental Microbiology* **17,**
790 2935-2951.
791 Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-
792 synonymous to synonymous rate ratio of all protein-coding genes in a genome. In: Peacock

C. (eds) Parasite Genomics Protocols. *Methods in Molecular Biology*, vol **1201,** Humana Press, New York, NY

Jones P, Binns D, Chang H-Y, *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236-1240.

Käll L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Research* **35,** W429-32.

Kellogg, EA (2001) Evolutionary history of the grasses. *Plant Physiology* **125,** 1198-1205.

Kind S, Hinsch J, Vrabka J, *et al.* (2018) Manipulation of cytokinin level in the ergot fungus *Claviceps purpurea* emphasizes its contribution to virulence. *Current Genetics* **64,** 1303-1319.

Kind S, Schurack S, Hinsch J, Tudzynski P (2018) *Brachypodium distachyon* as alternative model host system for the ergot fungus *Claviceps purpurea*. *Molecular Plant Pathology* **19,** 1005-1011.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **305,** 567-580.

Lee D, Jang E-H, Lee M, Kim S-W, Lee Y, Lee K-T, Bahn Y-S (2019) Unraveling melanin biosynthesis and signaling networks in *Cryptococcus neoformans*. *mBio* doi: 10.1128/mBio.02267-19.

Mahmoud A, Taga M (2012) Cytological karyotyping and characterization of a 410 kb minichromosome in *Nectria haematococca* MPI. *Mycologia* **104,** 845-56.

McCarthy CGP, Fitzpatrick DA (2019) Pan-genome analysis of model fungal species. *Microbial Genomics* **5,** e000243.

McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. *Nature Microbiology* **2,** 17040.

Menzies JG, Klein-Gebbinck HW, Gordon A, O'Sullivan DM (2017) Evaluation of *Claviceps purpurea* isolates on wheat reveals complex virulence and host susceptibility relationships. *Canadian Journal of Plant Pathology* **39,** 307-317.

Menzies JG, Turkington TK (2015) An overview of the ergot (*Claviceps purpurea*) issue in western Canada: challenges and solutions. *Canadian Journal of Plant Pathology* **37,** 40-51.

Mey G, Held K, Scheffer J, Tenberge KB, Tudzynski P (2002) CPMK2, an SLT2-homologous mitogen-activated protein (MAP) kinase, is essential for pathogenesis of *Claviceps purpurea* on rye: evidence for a second conserved pathogenesis-related MAP kinase cascade in phytopathogenic fungi. *Molecular Microbiology* **46,** 305-318.

Mey G, Oeser B, Lebrun MH, Tudzynski P (2001) The biotrophic, non-appressorium-forming grass pathogen *Claviceps purpurea* needs a *Fus3/Pmk1* homologous mitogen-activated protein kinase for colonization of rye ovarian tissue. *Molecular Plant-Microbe Interactions* **15,** 303-312.

Morran LT, Schmidt OG, Gelarden IA, Parrish RC 2nd, Lively CM (2011) Running with the Red Queen: host-parasite coevolution selects for biparental sex. *Science* **333,** 216-8

Negård M, Uhlig S, Kauserud H, Andersen T, Høiland K, Vrålstad T (2015) Links between genetic groups, indole alkaloid profiles and ecology within the grass-parasitic *Claviceps purpurea* species complex. *Toxins* **7,** 1431–1456.

Neubauer L, Dopstadt J, Humpf H-U, Tudzynski P (2016) Identification and characterization of the ergochrome gene cluster in the plant pathogenic fungus *Claviceps purpurea*. *Fungal Biology and Biotechnology* **3,** 2.

839 Nielsen H (2017) Predicting secretory proteins with SignalP In: *Protein function prediction* (eds.
840 Kihara D). *Methods in Molecular Biology* **1611,** Humana Press, New York, NY.
841 Oeser B, Heidrich PM, Muller U, Tudzynski P, Tenberge KB (2002) Polygalacturonase is a
842 pathogenicity factor in the *Claviceps purpurea*/rye interaction. *Fungal Genetics and Biology*
843 **36,** 176-186.
844 Oeser B, Kind S, Schurack S, Schmutzer T, Tudzynski P, Hinsch J (2017) Cross-talk of the
845 biotrophic pathogen *Claviceps purpurea* and its host *Secale cereale*. *BMC Genomics* **18,** 273.
846 Pažoutová S, Olšovská J, Linka M, Kolínská R, Flieger M (2000) Chemoraces and habitat
847 specialization of *Claviceps purpurea* populations. *Applied and Environmental Microbiology*
848 **66,** 5419–5425.
849 Pažoutová S, Pešicová K, Chudíčková M, Šrůtka P, Kolařík M (2015) Delimitation of cryptic
850 species inside *Claviceps purpurea*. *Fungal Biology* **119,** 7–26.
851 Pažoutová S, Raybould AF, Honzátko A, Kolínská R (2002) Specialised population of *Claviceps*
852 *purpurea* from salt marsh *Spartina* species. *Mycological Research* **106,** 210–214.
853 Peng Z, Oliveria-Garcia E, Lin G, *et al.* (2019) Effector gene reshuffling involves dispensable
854 mini-chromosomes in the wheat blast fungus. *PLoS Genetics* **15,** e1008272.
855 Píchová K, Pažoutová S, Kostovčík M, *et al.* (2018) Evolutionary history of ergot with a new
856 infrageneric classification (Hypocreales: Clavicipitaceae: *Claviceps*). *Molecular*
857 *Phylogenetics and Evolution* **123,** 73-87.
858 Pusztahelyi T, Holb IJ, Pócsi I (2015) Secondary metabolites in fungus-plant interactions.
859 *Frontiers in Plant Science* **6,** doi: 10.3389/fpls.2015.00573.

860 Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn R (2018) The MEROPS
861 database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison
862 with peptidases in the PANTHER database. *Nucleic Acids Research* **46,** 624-632.
863 Raybould AF, Gray AJ, Clarke RT (1998) The long-term epidemic of *Claviceps purpurea* on
864 *Spartina anglica* in Poole Harbour: pattern of infection, effects on seed production and the
865 role of *Fusarium heterosporum*. *New Phytologist* **138,** 497–505.
866 Rolke Y, Tudzynski P (2008) The small GTPase Rac and the p21-activated kinase Cla4 in
867 *Claviceps purpurea*: interaction and impact on polarity, development, and pathogenicity.
868 *Molecular Microbiology* **68,** 405-423.
869 Sánchez-Vallet A, Fouché S, Fudal I, Hartmann F, Soyer JL, Tellier A, Croll D (2018) The
870 genome biology of effector gene evolution in filamentous plant pathogens. *Annual Review of*
871 *Phytopathology* **56,** 21-40.
872 Sandve SR, Fjellheim S (2010) Did gene family expansions during the Eocene- Oligocene
873 boundary climate cooling play a role in Pooideae adaptation to cool climates? *Molecular*
874 *Ecology* **19,** 2075–2088.
875 Schardl CL, Young CA, Hesse U, *et al*. (2013) Plant-symbiotic fungi as chemical engineers:
876 multi-genome analysis of the Clavicipitaceae reveals dynamics of alkaloid loci. *PLoS*
877 *Genetics* **9,** e1003323.
878 Scheffer J, Chen C, Heidrich P, Dickman MB, Tudzynski P (2005a) A CDC42 homologue in
879 *Claviceps purpurea* is involved in vegetative differentiation and is essential for
880 pathogenicity. *Eukaryotic Cell* **4,** 1228-1238.

881 Scheffer J, Ziv C, Yarden O, Tudzynski P (2005b) The COT1 homologue CPCOT1 regulates
882 polar growth and branching and is essential for pathogenicity in *Claviceps purpurea*. *Fungal*
883 *Genetics and Biology* **42,** 107-118.
884 Shoukouhi P, Hicks C, Menzies JG, Popovic Z, Chen W, Seifert KA, Assabgui R, Liu M (2019)
885 Phylogeny of Canadian ergot fungi and a detection assay by real-time polymerase chain
886 reaction. *Mycologia* **111,** 1-13.
887 Smit AFA, Hubley R, Green P (2015) RepeatMasker Open-40. Available online at:
888 http://wwwrepeatmaskerorg.
889 Smit AFA, Hubley R, Green P (2015) RepeatMasker Open-40. Available online at:
890 http://wwwrepeatmaskerorg.
891 Song H, Nan Z, Song Q, Xia C, Li X, Yao X, Xu W, Kuang Y, Tian P, Zhang Q (2016)
892 Advances in research on *Epichloë* endophytes in Chinese native grasses. *Frontiers in*
893 *Microbiology* **7,** 1399.
894 Sperschneider J, Dobbs PN, Gardiner DM, Singh KB, Taylor JM (2018) Improved prediction of
895 fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology* **19,**
896 2094-2110.
897 Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of
898 large phylogenies. *Bioinformatics* **30,** 1312-1313.
899 Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification
900 of de novo predicted LTR retrotransposons. *Nucleic Acids Research* **37,** 7002–7013.
901 Stukenbrock EH, Dutheil JY (2018a) Fine-scale recombination maps of fungal plant pathogens
902 reveal dynamic recombination landscape and intragenic hotspots. *Genetics* **208,** 1209-1229.

903 Testa A, Oliver R, Hane J (2015) Overview of genomic and bioinformatic resources for
904 *Zymoseptoria tritici*. *Fungal Genetics and Biology* **79,** 13-16.
905 Tiley GP, Burleigh JG (2015) The relationship of recombination rate, genome structure, and
906 patterns of molecular evolution across angiosperms. *BMC Evolutionary Biology* **15,** 194.
907 Torriani SFF, Melichar JPE, Mills C, Pain N, Sierotzki H, Courbot M (2015) *Zymoseptoria*
908 *tritici*: a major threat to wheat production, integrated approaches to control. *Fungal Genetics*
909 *and Biology* **79,** 8-12.
910 Tudzynski P., Neubauer L. (2014) Ergot Alkaloids. In: Martín JF., García-Estrada C., Zeilinger
911 S. (eds) Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites. Fungal
912 Biology. Springer, New York, NY.
913 Urquhart AS, Mondo SJ, Makela MR *et al.* (2018) Genomic and genetic insights into a
914 cosmopolitan fungus, *Paecilomyces variotii* (Eurotiales). *Frontiers in Microbiology* **9,** 3058.
915 Van de Wouw A, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ
916 (2010) Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by
917 genomic environment and exposure to resistance genes in host plants. *PLoS Pathogens* **6,**
918 e1001180.
919 Wäli PP, Wäli PR, Saikkonen K, Tuomi J (2013) Is the pathogenic ergot fungus a conditional
920 defensive mutualist for its host grass? *PLoS ONE* **8,** e69249.
921 Wall JD, Stevison LS (2016) Detecting recombination hotspots from patterns of linkage
922 disequilibrium. *G3* **6,** 2265-2271.
923 Welch T, Feechan A, Kildea S (2018) Effect of host resistance on genetic structure of core and
924 accessory chromosomes in Irish *Zymoseptoria tritici* populations. *European journal of plant*
925 *pathology* **150,** 139-148.

926 Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs.
927 *Bioinformatics* **29,** 2487-2489.
928 Wyka SA, Mondo SJ, Liu M, Dettman J, Nalam V, Broders KD (2020a) Whole genome
929 comparisons of ergot fungi reveals the divergence and evolution of species within the genus
930 *Claviceps* are the result of varying mechanisms driving genome evolution and host range
931 expansion. bioRxiv: 10.1101/2020.04.13.039230.
932 Xia C, Christensen MJ, Zhang X, Nan Z (2018) Effect of *Epichloë gansuensis* endophyte
933 transgenerational effects on the water use efficiency, nutrient and biomass accumulation of
934 *Achnatherum inebrians* under soil water deficit. *Plant Soil* **424,** 555-571.
935 Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR
936 retrotransposons. *Nucleic Acids Research* **35,** W265–268.
937 Yang Z (2007) PAML4: phylogenetic analysis by maximum likelihood. *Molecular Biology and*
938 *Evolution* **24,** 1586-1591.
939 Yang Z, Neilsen R (2000) Estimating synonymous and nonsynonymous substitution rates under
940 realistic evolutionary models. *Molecular Biology and Evolution* **17,** 32-34.
941 Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance
942 computing toolset for relatedness and principal component analysis of SNP data.
943 *Bioinformatics* **28,** 3326-3328.
944
945 **Figure Captions:**
946
947 **Fig. 1. The pangenome of *Claviceps purpurea*.**

948 **(A)** Categorization of orthogroups (gene clusters) into core (shared between all isolates),
949 accessory (shared between ≥ 2 isolates, but not all), and singletons (found in only one isolate)
950 according to the number of orthrogroups shared between genomes. **(B)** Copy number variation in
951 core orthogroups containing paralogs. **(C)** Presence/absence variation and copy number variation
952 of accessory orthogroups, not including singletons. **(D)** Estimation of core and pangenome (core
953 + accessory + singleton) sizes by random resampling of possible combinations of 1 – 24
954 genomes (dots). Curves were modelled by fitting the power law regression formula: $y = Ax^B +$
955 C.
956
957 **Fig. 2. Analysis of predicted protein function across the *Claviceps purpurea* pangenome.**
958 Graphs indicate the proportion of orthogroups within each pangenome category of classified
959 protein function. **(A)** Containing conserved protein domains, **(B)** genes found in secondary (2°)
960 metabolite clusters, **(C)** possessing predicted secreted signals, **(D)** predicted to be effectors, **(E)**
961 containing transmembrane domains, **(F)** containing MEROPs domains for proteases and
962 peptidases, **(G)** contain CAZY enzymes, **(H)** all unclassified orthogroups not falling into a
963 previous category. Different letters (within each classification) represent significant differences
964 determined by multi-test corrected Fisher exact test ($P < 0.01$).
965
966 **Fig. 3. Distribution of omega (ω, dN/dS) ratios within the *Claviceps purpurea* core genome.**
967 Violin plots of ω ratios for core single-copy orthogroups protein functional categories. Solid
968 vertical lines within each plot represent the median, while dotted lines represent the 25th and 75th
969 quartile, respectively. Different letters represent significant differences determined by Kruskal-
970 Wallis with *post hoc* multi-test corrected Mann-Whitney U Test ($P \leq 0.01$).

971

972 **Fig. 4. Positive selection landscape within the *Claviceps purpurea* core genome.**

973 Positive selection of core single-copy orthogroups protein functional categories as predicted by
974 PAML with the CodeML algorithm. Genes with positive selection signatures were selected after
975 a stringent filtering around an $\alpha \leq 0.01$ (*See Methods*). **(A)** The total number of orthogroups in
976 functional categories with signatures of positive selection (outer circle). Omega (ω, dN/dS)
977 ratios of orthogroups within each functional category (inner circle). **(B)** The proportion of
978 orthogroups in each functional category based on the number of orthogroups examined in each
979 category. **(C)** The number of codons with selection signatures in the M8 model of CodeML, as
980 determined by the Bayes Empirical Bayes (BEB) algorithm with an $\alpha \leq 0.01$. Different letters **(B,**
981 **C)** represent significant differences determined by Kruskal-Wallis with *post hoc* multi-test
982 corrected Mann-Whitney U Test ($P \leq 0.01$). See Additional File 1 Fig. S5 for results from a less
983 stringent filtering of $\alpha \leq 0.05$.

984

985 **Fig. 5. Population recombination rates of representative scaffolds.**

986 Estimates of population recombination rates (ρ), in non-overlapping 1 kb windows, across four
987 representative scaffolds displaying the different variation observed across the *Claviceps*
988 *purpurea* genome. Smoothing curves were calculated from population recombination rates in 10
989 kb windows. See Additional File 1 Fig. S6 for remaining scaffolds.

990

991 **Fig. 6. Fine-scale recombination patterns across the *Claviceps purpurea* genome.**

992 Plots indicate the distribution of estimated population recombination rates (ρ) between **(A)**
993 different gene features (exons, introns, 500bp upstream and downstream), and **(B-D)** genes of

994 different functional categories and classification. Different letters represent significant
995 differences determined by Kruskal-Wallis with *post hoc* multi-test corrected Mann-Whitney U
996 Test ($P \leq 0.01$) between data within each plotting window, *** $P < 0.0001$. Sample sizes are
997 embedded below each plot.
998
999 **Fig. 7. Recombination hotspots predicted in *Claviceps purpurea* with associated genes and**
1000 **transposable elements (TEs).**
1001 Panels indicate scaffolds: **(A)** scaffold 14; **(B)** scaffold 15; **(C, D)** scaffold 20; **(E)** scaffold 23.
1002 Lines indicate background population recombination rates (ρ) estimated in non-overlapping 1 kb
1003 windows. Blue bars represent the position, intensity, and width of the predicted hotspots. Genes
1004 within the hotspot window and surrounding (± 20 kb) region are depicted by arrows with
1005 modified protein ID's of the reference (strain 20.1; append prefix of "CCE" and suffix of ".1" for
1006 protein ID's) from NCBI. Genes identified as duplicated (≥ 80% identity) from Wyka *et al.*
1007 2020a are outlined in red. TEs are depicted by lines between genes and the corresponding
1008 hotspot graph. Colors of arrows and lines correspond to the legend on the right.
1009
1010 **Supplemental Figure Captions:**
1011
1012 **Additional File 1 Fig S1. Genetic diversity of 24 *Claviceps purpurea* isolates.**
1013
1014 **Additional File 1 Fig S2. Average protein lengths (aa) of all orthogroups in *Claviceps***
1015 ***purpurea* pangenome.**
1016

Additional File 1 Fig S3. Distributions of mean non-synonymous (dN) and synonymous (dS) substitution rates of core single-copy orthogroups in *Claviceps purpurea*.

Additional File 1 Fig S4. Distributions of mean nucleotide identity (%) of core single-copy orthogroups in *Claviceps purpurea*.

Additional File 1 Fig S5. Positive selection landscape within the *Claviceps purpurea* core genome.

Additional File 1 Fig S6. Estimated population recombination rates of *Claviceps purpurea* scaffolds.

Additional File 1 Fig S7. Distributions of genes and their association (distance and flanking counts) to LTR transposable elements.

Additional File 1 Fig S8. Association of genes within recombination hotspots.

Additional File 1 Fig S9. Correlation of recombination rates and omega ratios.

Additional File 2 Table S1. *Claviceps purpurea* pangenome spreadsheet.

Additional File 3 Table S2. Enrichment of protein domains within pangenome.

Additional File 3 Table S3. Enrichment of protein domains within paralogous orthogroups.

Additional File 3 Table S4. PAML and CodeML summarized results.

Additional File 3 Table S5. BLAST results of single-copy core orthologs with an ω (dN/dS) ≥ 1.

Additional File 3 Table S6. Enrichment of protein domains of genes with positive selection.

Additional File 3 Table S7. Positive selection sites within five genes from two known *Claviceps purpurea* biosynthetic clusters.

Additional File 3 Table S8. Annotation information of missing reference scaffolds from 24 isolate whole-genome alignment.

Table 1: Collection and annotation statistics for the 24 *Claviceps purpurea* genomes used in this study.

Strain ID[†]	Origin	Host	Genome size (Mb)	Genomic GC (%)	TE[‡] content (%)	Gene count	BUSCO[§] score (%)
LM46	Canada: Alberta	*T. turgidum subsp. durum*	30.6	51.80%	9.64%	8,455	97.00%
LM60	Canada: Manitoba	*Avena sativa*	30.6	51.70%	9.29%	8,498	97.10%
LM223	Canada: Manitoba	*Bromus riparius*	30.8	51.70%	10.53%	8,438	96.60%
LM207	Canada: Manitoba	*Elymus repens*	30.5	51.80%	9.18%	8,475	97.00%
LM5	Canada: Manitoba	*Hordeum vulgare*	30.5	51.80%	8.95%	8,508	97.40%
LM33	Canada: Manitoba	*Hordeum vulgare*	30.5	51.80%	9.20%	8,557	97.10%
LM232	Canada: Manitoba	*Phalaris canariensis*	30.7	51.70%	9.36%	8,512	96.70%
LM233	Canada: Manitoba	*Phalaris canariensis*	30.6	51.80%	9.89%	8,717	96.60%
LM4	Canada: Manitoba	*Tricosecale*	30.6	51.80%	10.04%	8,470	96.90%
LM470	Canada: Ontario	*Elymus repens*	30.5	51.80%	8.95%	8,591	96.80%
LM474	Canada: Ontario	*Hordeum vulgare*	30.6	51.80%	9.38%	8,500	97.20%
LM469	Canada: Ontario	*Triticum aestivum*	30.5	51.80%	10.01%	8,394	96.50%
LM461	Canada: Quebec	*Elymus repens*	30.5	51.80%	8.42%	8,656	97.30%
LM14	Canada: Saskatchewan	*Hordeum vulgare*	30.6	51.80%	9.96%	8,422	97.30%
LM30	Canada: Saskatchewan	*Hordeum vulgare*	30.6	51.80%	9.35%	8,526	96.30%
LM39	Canada: Saskatchewan	*T. turgidum subsp. durum*	30.5	51.80%	10.11%	8,591	97.00%
LM28	Canada: Saskatchewan	*Triticum aestivum*	30.6	51.70%	9.58%	8,713	97.00%
LM582	Europe: Czech Republic	*Secale cereale*	30.7	51.80%	9.55%	8,518	95.50%
20.1	Europe: Germany	*Secale cereale*	32.1	51.60%	10.87%	8,703	95.50%
LM71	Europe: United Kingdom	*Alopercurus myosuroides*	30.5	51.80%	9.59%	8,472	97.00%
Clav55	Oceania: New Zealand	*Lolium perenne*	30.7	51.80%	9.80%	8,480	97.00%
Clav04	USA: Colorado	*Bromus inermis*	31.8	51.70%	10.05%	8,824	97.70%
Clav26	USA: Colorado	*Hordeum vulgare*	30.8	51.70%	9.07%	8,737	98.00%
Clav46	USA: Wyoming	*Secale cereale*	30.8	51.70%	9.68%	8,597	97.10%

† NCBI BioProject: PRJNA528707 (except 20.1, NCBI Accession = SAMEA2272775)

‡ Transposable element content presented in Wyka *et al.* 2020a, as a proportion of genomic sequences

§ Benchmarking Universal Single-Copy Orthologs Dikarya database (odb9)

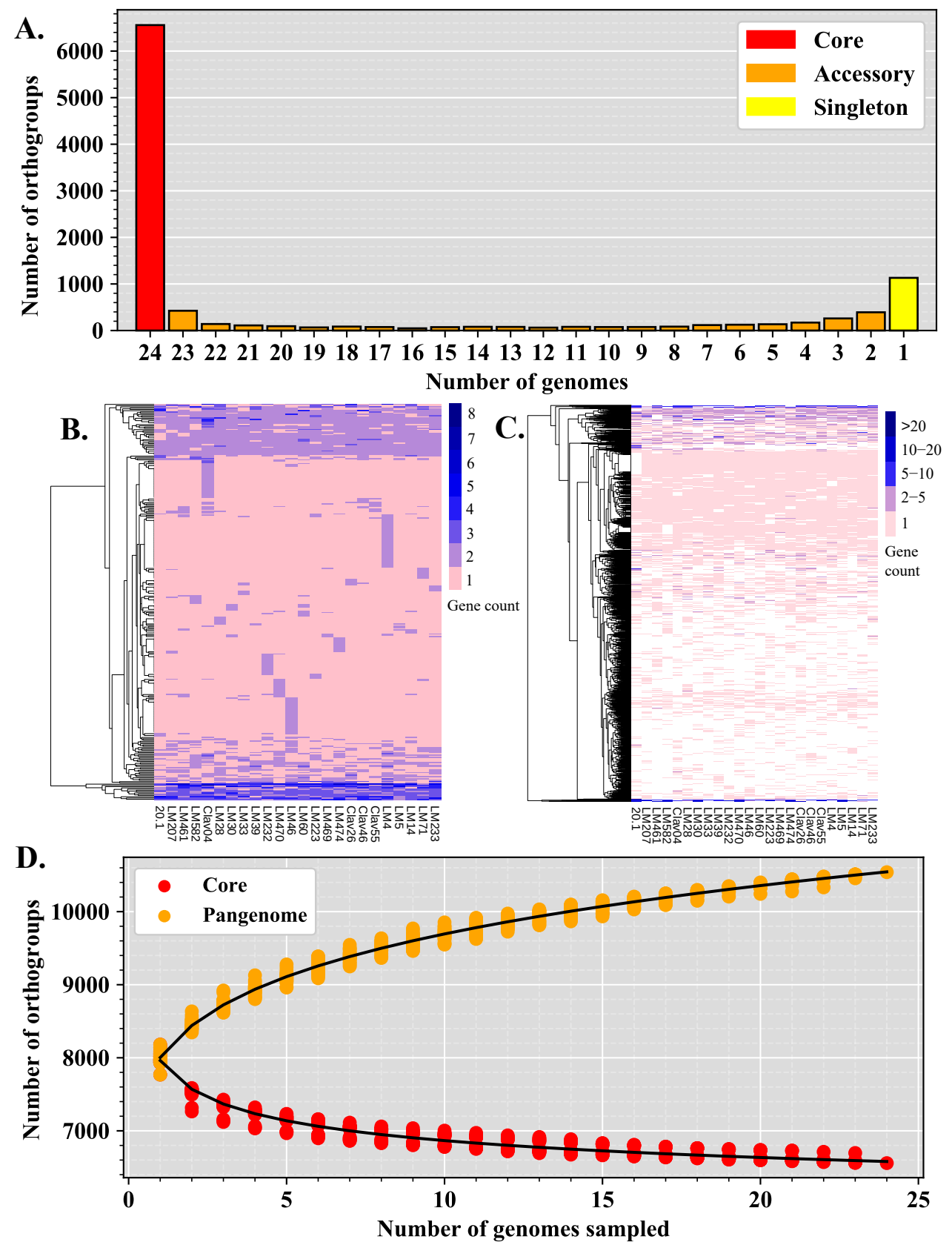
Table 2: PAML and CodeML processing information and filtering of core orthogroups for calculation of dN/dS (ω) ratios and examination of positive selection signatures.

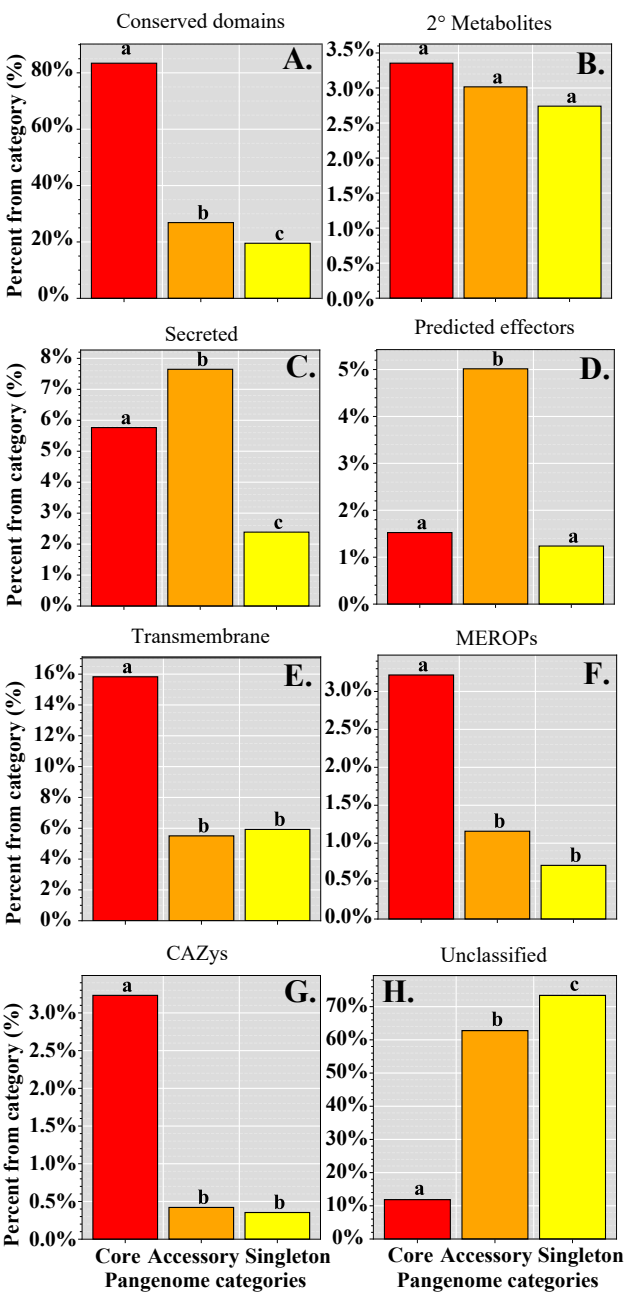
Total gene clusters (Pangenome)	10,540		
Single-copy gene clusters	6,244		
Number of clusters with N/A PAML results	43		
Cluster Classification (non-redundant) †:	Total Pangenome	Total Core‡	Single copy§
Effectors	257	100 (38.9%)	84 (84.0%)
Secreted	366	278 (75.9%)	253 (91.0%)
2° Metabolites	313	202 (64.5%)	181 (89.6%)
Transmembrane	1,210	998 (82.5%)	949 (95.1%)
MEROPs	167	149 (89.2%)	143 (96.0%)
CAZys	75	68 (90.7%)	66 (97.1%)
Conserved	4,754	3,985 (83.8%)	3,808 (95.6%)
Unclassified	3,398	778 (22.9%)	717 (92.2%)

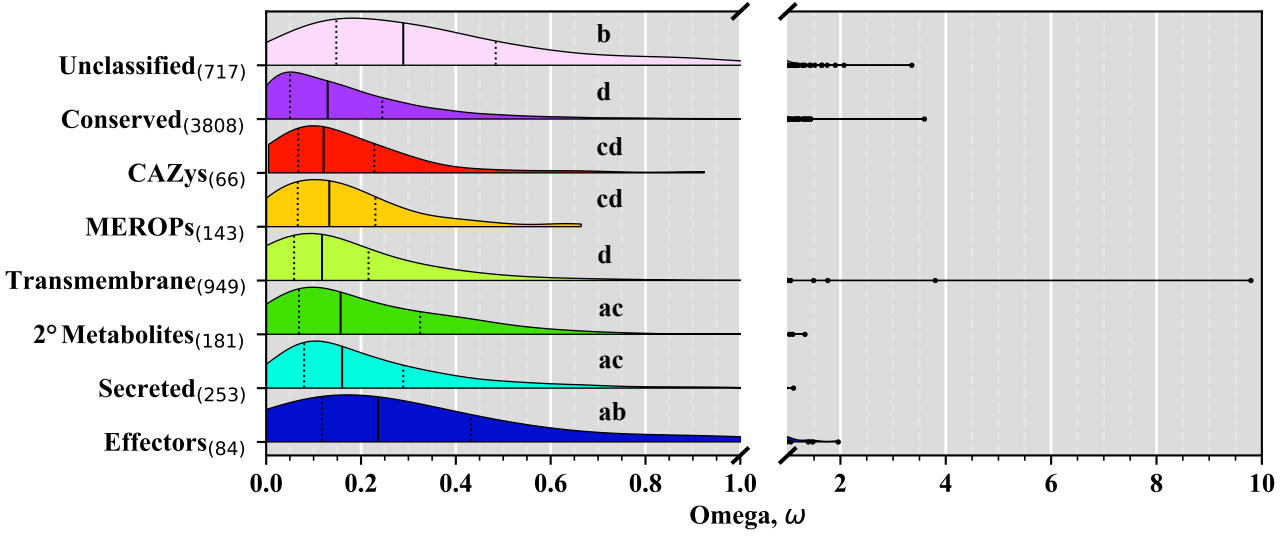
† For statistical purposes classification is structured such that each cluster is only represented once (in the order provided), i.e. secreted clusters are those not already classified as effectors, etc.
‡ Percentage out of total pangenome
§ Percentage out of total core

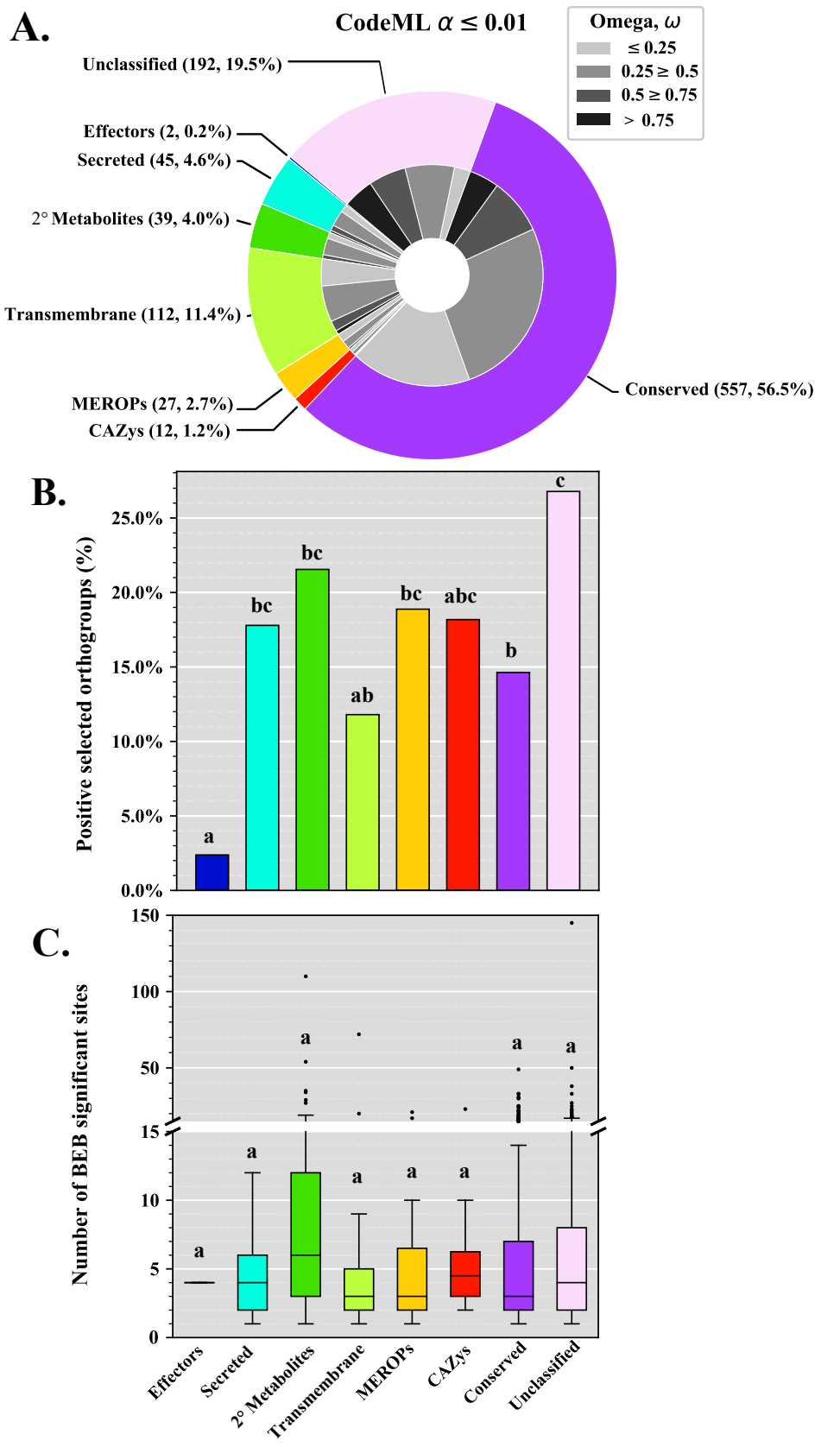
Table 3: Summary statistics of whole-genome alignment filtering and SNP calls for *Claviceps purpurea*.

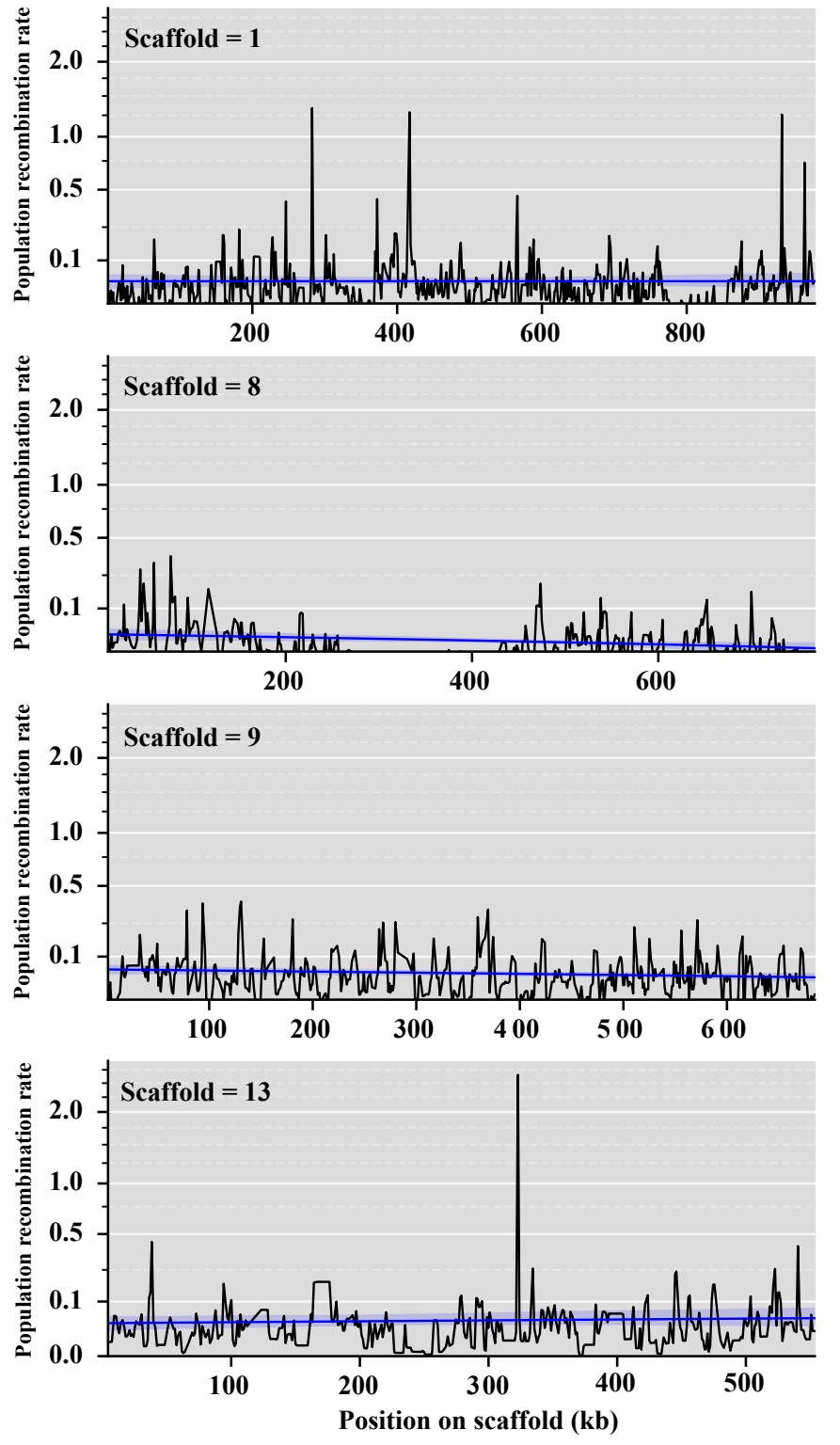
	C. purpurea strain 20.1	
Number of scaffolds	191	
Size of reference genome (bp)	32,091,443	
Number of exonic sites in reference genome (bp)	12,774,951 (39.8%)	
Number of haplotypes	24	
Summary Genome alignment:	Total Alignment Length (bp)	Number of alignment blocks
MultiZ alignment	27,523,755	16,330
Keep blocks with all strains	27,517,978	15,861
MAFFT in 10kb windows	27,378,024	15,870
Filter 1	26,198,304	57,891
Filter 2	24,959,120	97,532
Merged per contigs (N's filled in)	31,389,412	154
Total number of SNPs	1,152,999	
Total number of analyzed SNPs (biallelic, no unresolved state) and percent of total SNPs	1,076,901 (93.4%)	
Total number of SNPs in exons and percent of total	370,045 (32.1%)	
Total number of analyzed SNPs in exons (biallelic, no unresolved state) and percent of total analyzed SNPs in exons	358,258 (96.8%)	
Diversity in 10kb windows:	Median	
Watterson's Θ	0.01196	
Tajima's D	-0.82522	

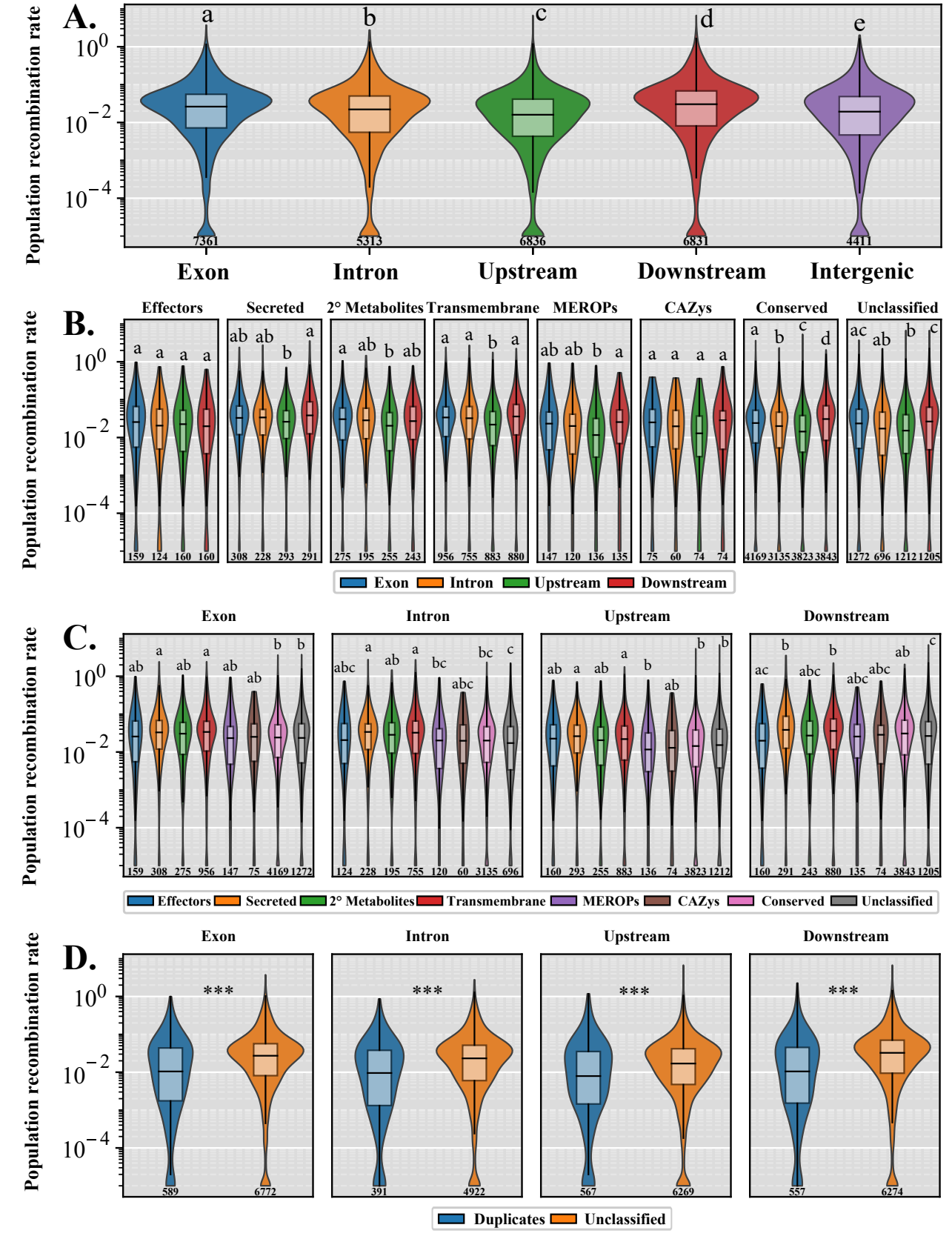

A.
Number of orthogroups
6000
5000
4000
3000
2000
1000
0
Core
Accessory
Singleton
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Number of genomes
B.
8
7
6
5
4
3
2
1
Gene count
20.1
LM207
LM461
LM582
Clav04
LM28
LM30
LM33
LM39
LM232
LM470
LM46
LM60
LM223
LM469
LM474
Clav26
Clav46
Clav55
LM4
LM5
LM14
LM71
LM233
C.
>20
10–20
5–10
2–5
1
Gene count
20.1
LM207
LM461
LM582
Clav04
LM28
LM30
LM33
LM39
LM232
LM470
LM46
LM60
LM223
LM469
LM474
Clav26
Clav46
Clav55
LM4
LM5
LM14
LM71
LM233
D.
Number of orthogroups
10000
9000
8000
7000
Core
Pangenome
0
5
10
15
20
25
Number of genomes sampled

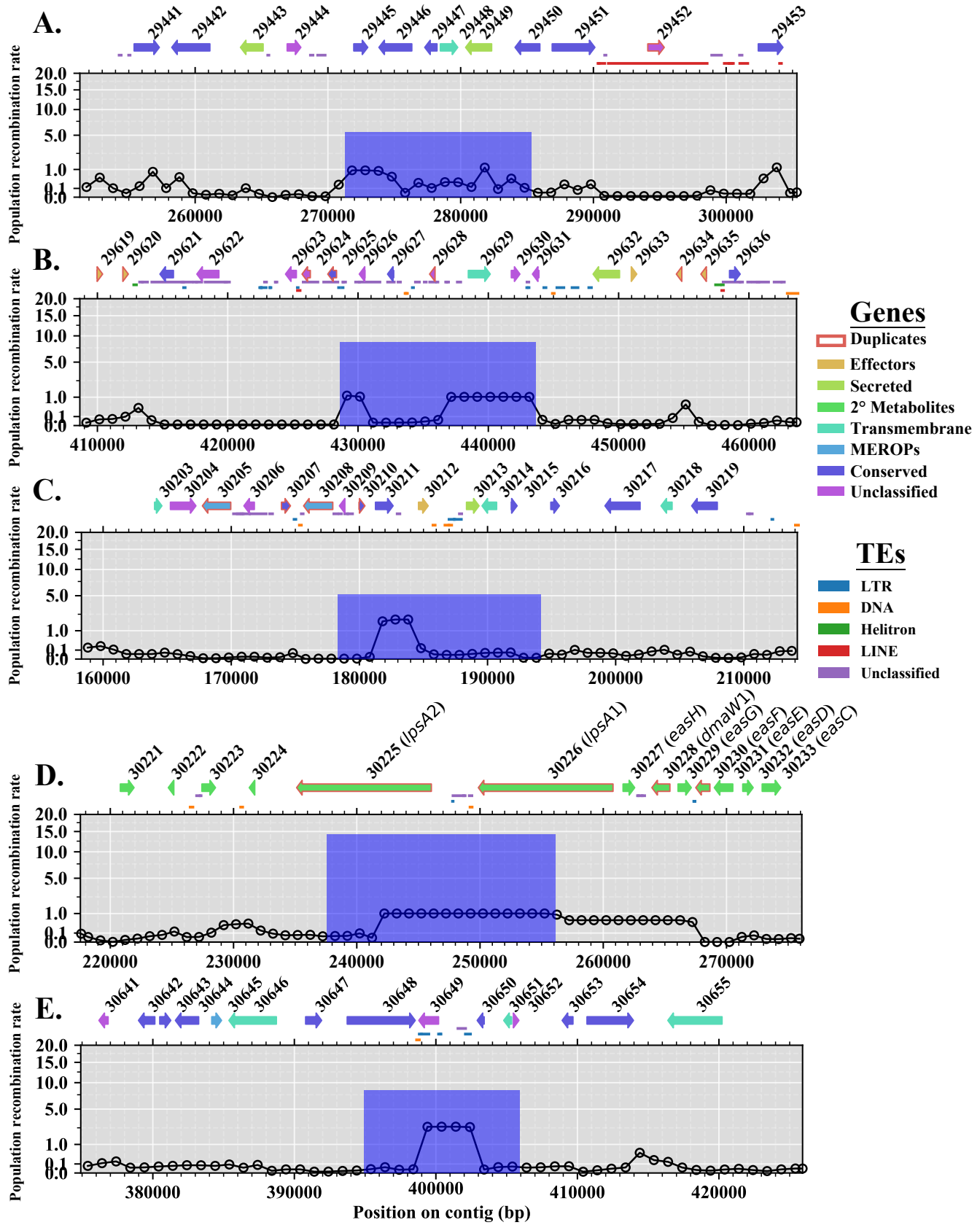

Conserved domains
A.
Percent from category (%)
80%
60%
40%
20%
0%
a
b
c
2° Metabolites
B.
3.5%
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
a
a
a
Secreted
C.
8%
7%
6%
5%
4%
3%
2%
1%
0%
a
b
c
Predicted effectors
D.
5%
4%
3%
2%
1%
0%
a
b
a
Transmembrane
E.
16%
14%
12%
10%
8%
6%
4%
2%
0%
a
b
b
MEROPs
F.
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
a
b
b
CAZys
G.
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
a
b
b
Unclassified
H.
70%
60%
50%
40%
30%
20%
10%
0%
a
b
c
Core Accessory Singleton
Pangenome categories
Core Accessory Singleton
Pangenome categories


Unclassified(717)
Conserved(3808)
CAZys(66)
MEROPs(143)
Transmembrane(949)
2° Metabolites(181)
Secreted(253)
Effectors(84)
b
d
cd
cd
d
ac
ac
ab
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Omega, ω


A.
CodeML $\alpha \leq 0.01$
Omega, ω
≤ 0.25
$0.25 \geq 0.5$
$0.5 \geq 0.75$
> 0.75
Unclassified (192, 19.5%)
Effectors (2, 0.2%)
Secreted (45, 4.6%)
2° Metabolites (39, 4.0%)
Transmembrane (112, 11.4%)
MEROPs (27, 2.7%)
CAZys (12, 1.2%)
Conserved (557, 56.5%)
B.
Positive selected orthogroups (%)
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
a
bc
bc
ab
bc
abc
b
c
C.
Number of BEB significant sites
150
100
50
15
10
5
0
a
a
a
a
a
a
a
a
Effectors
Secreted
2° Metabolites
Transmembrane
MEROPs
CAZys
Conserved
Unclassified


Scaffold = 1
Population recombination rate
2.0
1.0
0.5
0.1
200
400
600
800
Scaffold = 8
Population recombination rate
2.0
1.0
0.5
0.1
200
400
600
Scaffold = 9
Population recombination rate
2.0
1.0
0.5
0.1
100
200
300
4 00
5 00
6 00
Scaffold = 13
Population recombination rate
2.0
1.0
0.5
0.1
0.0
100
200
300
400
500
Position on scaffold (kb)


A.
Population recombination rate
Exon
Intron
Upstream
Downstream
Intergenic
B.
Effectors
Secreted
2° Metabolites
Transmembrane
MEROPs
CAZys
Conserved
Unclassified
C.
D.
***
Duplicates


A.
B.
C.
D.
E.
Population recombination rate
20.0
15.0
10.0
5.0
1.0
0.1
0.0
29441
29442
29443
29444
29445
29446
29447
29448
29449
29450
29451
29452
29453
260000
270000
280000
290000
300000
29619
29620
29621
29622
29623
29624
29625
29626
29627
29628
29629
29630
29631
29632
29633
29634
29635
29636
410000
420000
430000
440000
450000
460000
30203
30204
30205
30206
30207
30208
30209
30210
30211
30212
30213
30214
30215
30216
30217
30218
30219
160000
170000
180000
190000
200000
210000
30221
30222
30223
30224
30225 (lpsA2)
30226 (lpsA1)
30227 (easH)
30228 (dmaW1)
30229 (easG)
30230 (easF)
30231 (easE)
30232 (easD)
30233 (easC)
220000
230000
240000
250000
260000
270000
30641
30642
30643
30644
30645
30646
30647
30648
30649
30650
30651
30652
30653
30654
30655
380000
390000
400000
410000
420000
Position on contig (bp)
Genes
Duplicates
Effectors
Secreted
2° Metabolites
Transmembrane
MEROPs
Conserved
Unclassified
TEs
LTR
DNA
Helitron
LINE
Unclassified