# Genomewide signatures of selection in Epichloë reveal candidate genes for host specialization

**Journal Article**

**Author(s):**
Schirrmann, Melanie K.; Zoller, Stefan; Croll, Daniel; Stukenbrock, Eva H.; Leuchtmann, Adrian; Fior, Simone

**Publication date:**
2018-08

**Permanent link:**
https://doi.org/10.3929/ethz-b-000281177



**Originally published in:**
Molecular Ecology 27(15), https://doi.org/10.1111/mec.14585

**Funding acknowledgement:**
138479 - Mechanisms of pre- and postzygotic isolation in cryptic Epichloë species (SNF)

# Genome-wide signatures of selection in *Epichloë* reveal candidate genes for host specialization

Melanie K. Schirrmann[1,2], Stefan Zoller[3], Daniel Croll[4], Eva. H. Stukenbrock[5], Adrian Leuchtmann[1] and Simone Fior[1]

[1]Institute of Integrative Biology (IBZ), ETH Zurich, Zürich, Switzerland

[2]Research Group Molecular Diagnostics, Genomics and Bioinformatics, Agroscope, Waedenswil, Switzerland

[3]Genetic Diversity Centre (GDC), ETH Zurich, Zürich, Switzerland

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

[5]Environmental Genomics, Christian-Albrechts University of Kiel, Kiel, Germany and Max Planck Institute for Evolutionary Biology, Plön, Germany.

**Corresponding author**

Melanie Katharina Schirrmann

Schloss 1

8820 Waedenswil

Email: m.schirrmann@live.de

Tel.: +41 58 46 99007

**Running title**

Host specialization candidates in *Epichloë*

**Keywords**

Endophytic fungi, host specialization, pathogens, population genomics, positive selection, secreted proteins

**Abstract**

Host specialization is a key process in ecological divergence and speciation of plant-associated fungi. The underlying determinants of host specialization are generally poorly understood, especially in endophytes, which constitute one of the most abundant components of the plant microbiome. We addressed the genetic basis of host specialization in two sympatric subspecies of grass-endophytic fungi from the *Epichloë typhina* complex; subsp. *typhina* and *clarkii*. The life cycle of these fungi entails unrestricted dispersal of gametes and sexual reproduction before infection of a new host, implying that the host imposes a selective barrier on viability of the progeny. We aimed to detect genes under divergent selection between subspecies, experiencing restricted gene flow due to adaptation to different hosts. Using pooled whole-genome sequencing data, we combined $F_{ST}$ and $D_{XY}$ population statistics in genome scans and detected 57 outlier genes showing strong differentiation between the two subspecies. Genome-wide analyses of nucleotide diversity ($\pi$), Tajima's $D$, and $d$N/$d$S ratios indicated that these genes have evolved under positive selection. Genes encoding secreted proteins were enriched among the genes showing evidence of positive selection, suggesting that molecular plant-fungus interactions are strong drivers of endophyte divergence. We focused on five genes encoding secreted proteins, which were further sequenced in 28 additional isolates collected across Europe to assess genetic variation in a larger sample size. Signature of positive selection in these isolates and putative identification of pathogenic function supports

51 our findings that these genes represent strong candidates for host specialization genes
52 in *Epichloë* endophytes. Our results highlight the role of secreted proteins as key
53 determinants of host specialization.

54 **Introduction**

55 Ecological divergence is a process whereby natural selection drives adaptation of
56 populations to distinct ecological environments (Arnegard *et al.* 2014). The genetic
57 architecture and underlying function of adaptive traits is at the core of evolutionary
58 biology studies aiming to understand how natural selection can lead to lineage
59 divergence and speciation. In recent years, genome scans have provided
60 unprecedented insights into the genetic determinants of ecological divergence in a
61 number of organisms, including well-studied model systems in evolutionary biology
62 such as sticklebacks (Jones *et al.* 2012), cichlid fishes (Brawand *et al.* 2014),
63 flycatchers (Ellegren *et al.* 2012), stick insects (Soria-Carrasco *et al.* 2014), hooded
64 crows (Poelstra *et al.* 2014), and *Heliconius* butterflies (Martin *et al.* 2013). One
65 underlying principle of genome scan based studies is that loci under divergent
66 selection experience less introgression compared to the rest of the genome as a
67 consequence of selection acting on these loci. This "protection" from the
68 homogenising effect of gene flow enables the formation of highly differentiated
69 regions (Wu 2001; Nosil *et al.* 2005; Feder *et al.* 2012).

70 One of the most widespread forms of ecological divergence in plant-associated
71 fungi is host-driven specialization (Vialle *et al.* 2013; Restrepo *et al.* 2014). The close
72 association between a symbiotic fungus and its host depends on the co-evolution of
73 physiological and life history traits linked to the interaction between the co-existing
74 organisms. Fungal infection of host plants is mediated by multiple signalling
75 events, including the secretion of proteins (often in the form of small secreted
76 proteins, so called effectors) that suppress immune responses or manipulate host cell
77 physiology in pathogenic systems (Rep 2005; Plissonneau *et al.* 2017). After
78 successful infection, pathogens exploit their host plants for nutrients to sustain their

79 own growth and reproduction while inducing disease and compromising host viability
80 and reproduction (Bronstein 2009). This antagonistic relationship of pathogens results
81 in a co-evolutionary arms race (Dawkins & Krebs 1979) by which genes involved in
82 the host-fungus interaction are continuously subject to selection in response to
83 changes occurring in the symbiotic partner (Presti *et al.* 2015). In particular, secreted
84 proteins that directly interact with host molecules are expected to be strong targets of
85 natural selection (Terauchi & Yoshida 2010). This expectation was confirmed by the
86 identification of signatures indicating positive selection on genes encoding secreted
87 proteins in a number of plant pathogens (Win *et al.* 2007; Barrett *et al.* 2009; Poppe *et*
88 *al.* 2015).

89 Genes encoding secreted proteins and other genetic determinants of host
90 specialization may also play a central role in speciation of plant pathogens (Giraud *et*
91 *al.* 2006; Giraud 2006). Some fungal plant pathogens are obligate biotrophs that
92 complete their entire life cycle on a single compatible host plant, and undertake sexual
93 reproduction within the plant without effective dispersal of gametes. As the ability to
94 infect a host depends on the necessary repertoire of effector proteins, mating partners
95 are determined by the set of effectors that allow infection of the same host (Giraud *et*
96 *al.* 2010). Host specialization can thus form a strong postzygotic barrier preventing
97 genome-wide introgression between strains specialized to distinct hosts (e.g. different
98 host adapted races), leading to species formation (Giraud *et al.* 2006). Other
99 pathogens have a life cycle that entails free movement of gametes mediated by wind
100 or a vector before mating takes place outside (e.g. on the ground) or on the surface of
101 the plant. After zygote formation and meiosis, haploid spores are dispersed and new
102 infections are determined by the ability of the progeny to infect the plant on which
103 spores have landed. In the absence of intrinsic prezygotic barriers and assortative

104 mating, selection on host specialization loci cannot impede exchange of neutral
105 genomic regions between pathogen races. Selection will maintain host specialization
106 alleles in each race, and reproductive isolation only evolves as a consequence of
107 assortative mating or active host choice (Giraud *et al.* 2006; Giraud 2006).

108 Genome scans can be used to detect outlier regions in the genome that stand
109 out with respect to the distribution of genetic variants. Different test statistics can be
110 used to identify regions with either an increased or reduced nucleotide differentiation.
111 Regions that have been subjected to divergent selection during ecological divergence
112 of two populations can typically be recognized by an increased differentiation using
113 $F_{ST}$-based statistics (Ellison *et al.* 2011; Branco *et al.* 2015). However, other
114 processes not necessarily related to divergent ecological specialization can produce
115 similar signatures in the genome sequence. Recent theoretical and empirical work
116 have highlighted the role of linked selection in generating a signature of increased
117 differentiation, especially in regions of low recombination (Cutter & Payseur 2013;
118 Wolf & Ellegren 2016). Furthermore, demographic history, evolutionary rates and
119 genomic architecture can affect the distribution of nucleotide differentiation (Vijay *et*
120 *al.* 2016; Van Doren *et al.* 2017), calling for awareness on the methodology employed
121 to associate highly differentiated regions with loci experiencing divergent selection
122 (Burri 2017).

123 At advanced stages of lineage separation, net divergence given by the
124 parameter $D_{XY}$ (Nei 1987) is expected to capture the level of polymorphism
125 accumulated since the divergence of populations. This parameter is suitable to infer
126 variation in the rate of gene flow across the genome, while sensitive to signatures of
127 selection in the ancestral population (Cruickshank & Hahn 2014; Guerrero & Hahn
128 2017). Loci underlying divergent selection are thus expected to show both high $F_{ST}$

129 and $D_{XY}$ values between ecologically diverging populations or species (Nachman &
130 Payseur 2012; Cruickshank & Hahn 2014). Additional test statistics to assess the
131 impact of natural selection on sequence evolution include measures of nucleotide
132 diversity ($\pi$), site frequency analyses (e.g. Tajima's *D*), and estimates of non-
133 synonymous and synonymous variation in coding sequences within and between
134 species. Combined evidence from these measures can strengthen inferences of
135 deviation from neutral evolution for putative adaptive loci (Wolf & Ellegren 2016).

136 Population genomics of host-specialized fungi have provided a powerful
137 approach to identify genes that have been under selection during the divergence of
138 populations, and that may have allowed the colonization of distinct hosts. Such
139 analyses have been used to detect genes involved in divergent host specialization in
140 the fungal pathogen species *Zymoseptoria tritici* (synonym *Mycosphaerella*
141 *graminicola*) and *Microbotryum lychnidis-dioicae* (Stukenbrock *et al.* 2011; Poppe *et*
142 *al.* 2015; Badouin *et al.* 2017). Both species establish intercellular networks that
143 resemble endophytic growth within the host tissues after successful infection. While *Z.*
144 *tritici* eventually switches to a necrotrophic growth after a long latent period, *M.*
145 *lychnidis-dioicae* sterilises its host for its own reproduction without killing it.
146 However, the genetic basis of host specialization remains poorly understood and more
147 studies are needed to dissect the underlying mechanisms of lineage divergence and
148 host specialization and to identify key determinants of symbiotic interactions.

149 *Epichloë* (Ascomycota, Clavicipitaceae) belongs to the large group of fungal
150 endophytes, one of the most diverse and abundant components of the plant
151 microbiome (Ganley *et al.* 2004; Busby *et al.* 2016). Sexual species of this genus
152 grow symptomless within plant tissues with no clear sign of defence response from
153 the plant during the plant vegetative phase (Schardl *et al.* 2004), but they severely

154 affect the plant inflorescence during the plant reproductive phase, which coincides
155 with the sexual stage of the fungal life cycle (Fig. 1; Leuchtmann & Schardl 1998). At
156 this time, the haploid fungal mycelium proliferates massively within the expanding
157 grass inflorescence to produce external fruiting structures (i.e. stromata) including
158 both male gametes (i.e. spermatia) and corresponding female receptive hyphae (White
159 *et al.* 1997). The reproduction of the fungus finally results in the sterilisation of host
160 flowering stems, causing a syndrome known as 'choke disease' (Western & Cavett
161 1959; Kirby 1961). *Epichloë* species are heterothallic (i.e. different mating types
162 prevent fertilization between spermatia and female structures from the same stroma),
163 and host plants are infected with only one strain, thus obligate outcrossing occurs
164 between genotypes having infected different plants. After successful mating,
165 karyogamy and meiosis take place on the stroma and haploid ascospores are wind-
166 dispersed and mediate horizontal transmission to new hosts by infection of grass
167 florets and seeds (Fig. 1).

168 In previous work, we focused on two sympatrically growing subspecies of
169 sexually reproducing *E. typhina* subsp. *typhina* infecting *Dactylis glomerata* and *E.*
170 *typhina* subsp. *clarkii* infecting *Holcus lanatus* (hereafter *E.t. typhina* and *E.t. clarkii*).
171 We found clear genotypic differentiation between the two subspecies (Schirrmann *et*
172 *al.* 2015), and reciprocal infections with host-associated strains showed host
173 specificity (Schirrmann & Leuchtmann 2015). Subspecies within the same species
174 complex can be crossed in artificial experiments (Leuchtmann & Schardl 1998), and
175 hybrids are viable *in vivo* following infection of parental as well as extra-parental host
176 plants (Schirrmann & Leuchtmann 2015). In natural ecosystems, mating is vectored
177 by non-selective flies of the genus *Botanophila* (Anthomyiidae) in a process similar to
178 pollination (Bultman *et al.* 1998), with potential hybridization of fungal subspecies

179 occurring in geographic proximity. Indeed, hybrid ascospores between *E.t. typhina*
180 and *E.t. clarkii* have previously been identified (Bultman *et al.* 2011). The life cycle
181 of *E.t. typhina* and *E.t. clarkii* conforms to a model of sexual reproduction where
182 mating can occur between individuals specialized to different host plants. More
183 specifically, there are effectively no intrinsic pre- and post-zygotic barriers to hybrid
184 formation, and selection imposed by host specialization may thus be the key
185 determinant of the ability of hybrid spore genotypes to infect a new host and
186 reproduce successfully. Following the classical model of lineage divergence occurring
187 in the presence of gene flow, strong allelic differentiation is expected at loci
188 underlying host specialization in contrast to the rest of the genome.

189 In this study, we aimed to identify candidate genes underlying host
190 specialization of *E.t. typhina* and *E.t. clarkii. Epichloë* fungi interact with the host
191 grass throughout its entire life cycle to establish and maintain infection. Given the
192 strict host specificity of the studied subspecies, we expected to find signatures of
193 divergent selection (i.e. selection acting in different directions on the two subspecies)
194 on genes encoding secreted proteins, as these may be involved in the specific
195 interaction with host molecules, as it has been shown in other pathogenic fungi (e.g.
196 Rep 2005; Terauchi & Yoshida 2010; Presti *et al.* 2015; Poppe *et al.* 2015; Badouin *et*
197 *al.* 2017). We analysed whole-genome pooled sequencing data from the two
198 sympatrically growing subspecies to detect outlier loci with signatures of increased
199 divergence. Our approach combined analyses of nucleotide differentiation based on
200 $F_{ST}$ and $D_{XY}$ to detect divergent selection between the two subspecies, and neutrality
201 tests (i.e. nucleotide diversity $\pi$ and Tajima's $D$) to detect deviations from neutrality
202 within the two subspecies. Furthermore, we inferred gene-wise estimates of non-
203 synonymous and synonymous divergence and polymorphisms to compute $d$N/$d$S

204 ratios between subspecies and to perform a McDonald-Kreitman (MK) test within
205 subspecies to detect signatures of positive selection (McDonald & Kreitman 1991;
206 Goldman & Yang 1994). By combining the outcome of selection scans with
207 functional gene predictions, we identified five candidate genes encoding secreted
208 proteins for which we confirmed signatures of positive selection using additional 28
209 isolates from European populations. We consider these five genes strong candidates
210 for host specialization determinants.
211

212 **Methods**

213 ***Population genomics sequencing***

214 A total of twenty haploid *E.t. typhina* stromata (i.e. fruiting structures) and twenty
215 haploid *E.t. clarkii* stromata, respectively, were sampled in spring 2013 from
216 sympatric populations at Aubonne, Switzerland. Individual stromata were collected
217 from infected plants spaced every five meters along eight transects. Given that
218 stromata contain both fungal and grass material, the interior part of each stroma was
219 split open under sterile conditions to separate mycelium from visible grass tissues.
220 Mycelial DNA was extracted using the DNeasy Plant Kit (Quiagen, Germantown,
221 MD, USA). DNA quality was checked on 1.5% agarose gels stained with GelRed
222 using a UV-Vis Spectrometer and DNA quantity was measured with a Qubit
223 fluorometer using the broad-range dsDNA standard. The population genomic dataset
224 was obtained using a pool sequencing (Pool-Seq) approach (Schlötterer *et al.* 2014).
225 High-quality DNA from mycelium of stromata of each subspecies was pooled in
226 equimolar amounts, producing one 5 μg RNA-free genomic DNA sample for each of
227 the subspecies. Illumina libraries of ~600 bp insert size were generated following the
228 instructions of the Illumina Paired-End Sample Preparation Kit. Sequencing was

229 performed on an Illumina MiSeq lane using 150 bp paired-end reads to produce an
230 expected coverage of ~150 X for each subspecies and an expected coverage of ~7.5 X
231 per individual. Reads have been submitted to the NCBI Sequence Read Archive
232 (SRA) under accession numbers SRR5571977 - SRR5571978.
233
234 ***Illumina read mapping and SNP calling***
235 To filter sequence reads, we used Trimmomatic (Bolger *et al.* 2014) to remove
236 Illumina adapters, bases at the start and end of a read below a quality threshold of 5,
237 and low-quality segments from the end of a read using a 4 bp sliding window and
238 threshold for average quality of 15. Trimmed reads shorter than 50 bp were discarded.
239 Remaining reads were mapped to an existing *E. typhina* subsp. *poae* genome
240 assembly (*E.t. poae*; E5819; http://www.endophyte.uky.edu; Schardl *et al.* 2013). *E.t.*
241 *poae* is a close relative of the studied subspecies and belongs to the same species
242 complex (Leuchtmann & Schardl 1998; Craven *et al.* 2001). The *E.t. poae* reference
243 genome includes 34 Mb assembled in 2072 contigs, with an N50 value of 36475 bp
244 (Schardl *et al.* 2013). In total, genic regions (including UTRs, exons and introns)
245 comprise 15.2 Mb (44.7%), coding sequences (exons only) compromise 10.5 Mb
246 (30.9%), and repetitive DNA compromises 41.6% (Schardl *et al.* 2013). Reads were
247 mapped with BWA-MEM version 0.7.8 using the default settings (Li & Durbin 2009).
248 Alignments were filtered for a minimum mapping quality of 20, and remaining high-
249 quality reads were sorted and indexed with Samtools v. 0.1.18
250 (http://samtools.sourceforge.net/). Single nucleotide polymorphisms (SNPs) within
251 each of the two host-associated subspecies were called with Samtools (mpileup; Li *et*
252 *al.* 2009) using default settings, and population statistic measures were computed
253 using software specifically developed for Pool-Seq data, i.e. PoPoolation (Kofler *et*

254 *al.* 2011a) and PoPoolation2 (Kofler *et al.* 2011b).

255

256 ***Population genomics analyses***

257 Given the higher proportion of reads mapping to coding regions of the *E.t. poae*
258 reference genome (see Results), all population genomic analyses were performed on
259 gene coding sequences. To detect candidate genes involved in host specialization, our
260 approach aimed to identify loci showing elevated differentiation and divergence as
261 inferred from $F_{ST}$ and $D_{XY}$ statistics, respectively. As a relative measure of
262 differentiation, $F_{ST}$ is sensitive to variation in within-population genetic diversity, and
263 heterogeneous patterns across the genome can arise from processes unrelated to host
264 specialization. $D_{XY}$ measures the average number of nucleotide differences between
265 populations, and is expected to be highest in genomic regions protected from gene
266 flow. Because sorting of ancestral variation and new mutations must occur in the
267 diverging populations for $D_{XY}$ to increase, a longer divergence time compared to $F_{ST}$
268 is required before a significant signal arises. Therefore, a signal is only detected at
269 advanced stages of divergence compared to $F_{ST}$. Importantly, the signals from $F_{ST}$ and
270 $D_{XY}$ are expected to overlap for loci under divergent selection that experience reduced
271 gene flow (Cruickshank & Hahn 2014). Inference of differentiation across the aligned
272 *E.t. typhina* and *E.t. clarkii* sequences was performed by computing $F_{ST}$ following the
273 approach of Hartl & Clark (2007) given by the formula $F_{ST} = (\pi_T - \pi_S)/\pi_T$, where $\pi_T$
274 is the expected heterozygosity in the total sample and $\pi_S$ is the expected average
275 heterozygosity in each population. Inference of divergence was calculated as $D_{XY}$ =
276 $\sum x_i y_i d_{ij}$, where, $d_{ij}$ measures the number of nucleotide differences between the $i^{th}$
277 haplotype from population $X$ and the $j^{th}$ haplotype from population $Y$ (Cruickshank &
278 Hahn 2014).

279 Gene-wise $F_{ST}$ values were calculated using Popoolation2 (Kofler *et al.*
280 2011b), which allows comparison of allelic SNP frequencies between two or more
281 populations. Scripts mentioned below are included in this package unless otherwise
282 specified. Synchronisation and filtering of the Samtools mpileup-file were performed
283 using *mpileup2sync.jar*. The gene annotation of *E.t. poae* was transferred onto the
284 synchronized file using *create-genewise-sync.pl* to generate a gene-based dataset for
285 population statistics. $F_{ST}$ estimates were obtained with *fst-sliding.pl*, setting the
286 number of chromosomes pooled per population to 20 and the window size (40,000
287 bp) longer than the length of any *E.t. poae* gene, as recommended by the authors of
288 the program (Kofler *et al.* 2011b). The minor allele count was set to two for each
289 gene, and at least 50% of SNPs had to fulfil the minimum coverage of six and
290 maximum coverage of 60 in each subspecies to minimise the risk of calling variants in
291 poorly mapped or repetitive regions. Allele frequencies based on SNPs were
292 estimated with *snp-frequency-diff.pl*, filtering with the same thresholds as described
293 for $F_{ST}$ estimates.

294 We computed the measure of divergence $D_{XY}$ for bi-allelic loci (~99.2% of the
295 total allele estimates) as the average number of nucleotide differences for each gene
296 based on the formula using allele frequencies from Smith & Kronforst (2013) defined
297 as $D_{XY} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ix}\left(1-\hat{p}_{iy}\right) + \hat{p}iy(1-\hat{p}_{ix})$. Genes with values above the 95%
298 quantile in both $F_{ST}$ and $D_{XY}$ population statistics were defined as outliers ($F_{ST}$-$D_{XY}$
299 outliers) and considered as candidate genes involved in host specialization. To obtain
300 further estimates of sequence diversity to confirm signatures of divergent selection for
301 the $F_{ST}$-$D_{XY}$ outliers we calculated nucleotide diversity π (Nei & Li 1979) and
302 Tajima's *D* (Tajima 1989). Reduced levels of nucleotide diversity and Tajima's *D* at
303 candidate genes conforms to expectations of positive selection acting on these loci. In

304 particular, Tajima's $D$ is 0 in a population evolving with mutation-drift equilibrium,
305 Tajima's $D > 1$ is indicative of balancing selection or population contraction, and
306 Tajima's $D < 1$ is indicative of recent positive selection or population expansion.
307 Although similar patterns in diversity measures can be generated by different
308 demographic processes, evidence from these statistics can be informative to support
309 departure from neutrality. Within both subspecies, gene-wise $\pi$ (Nei & Li 1979) and
310 Tajima's $D$ were computed using *variance-at-position.pl* in PoPoolation (Kofler *et al.*
311 2011a). Minimum requirements for coverage and allele count used in SNP calling
312 were set as described above for the $F_{ST}$ calculations. All statistical analyses were
313 conducted in R version 2.13.0 (R Development Core Team, 2011).
314
315 ***Prediction of gene functions***
316 The reference genome of *E.t. poae* (Schardl *et al.* 2013) includes a structural
317 annotation of genic regions, but a functional annotation is currently not available. We
318 produced a *de novo* functional annotation for *E.t. poae* using Blast2Go with a
319 similarity search against a local installation of the NCBI non-redundant (nr) database
320 (Conesa *et al.* 2005). A list of the functional annotations for all genes is available as
321 supplementary material (Table S1, Supporting information). In addition, protein
322 domain detection was performed with a local installation of InterProScan v.5RC7
323 (Jones *et al.* 2014).
324 To identify the biological processes associated with the genes identified as
325 candidates for host specialization among the $F_{ST}$-$D_{XY}$ outliers, we performed a Gene
326 Ontology (GO) enrichment analysis using Blast2Go (Conesa & Götz 2007).
327 Significance of each individual GO category was computed using a Fisher's exact test
328 with a significance threshold of 1%. Correction for multiple testing was performed

329 using a false discovery rate (FDR) of 0.05. The *E.t. poae* genes used for the $F_{ST}$
330 analyses were used as background reference for the analyses in *E.t. typhina* and *E.t.*
331 *clarkii*. Given that genes encoding secreted proteins may be involved in host-fungus
332 interactions, we screened predicted extra-cellular protein sequences for the presence
333 of a secretion signal, and transmembrane, cytoplasmic and extracellular domains
334 using a combination of SignalP v.4.1 (Petersen *et al.* 2011), Phobius v.1.01 (Käll *et*
335 *al.* 2004) and TMHMM v.2.0 (Krogh *et al.* 2001). Protein sequences of genes
336 included in the final candidate selection were further characterized with InterProScan
337 v.5.16-55.0 (Jones *et al.* 2014). To this end, we assigned protein sequence motifs to
338 protein families (PFAM) and GO categories based on hidden Markov models (HMM)
339 implemented in InterProScan.
340
341 ***Detection of orthologous and paralogous genes***
342 Genes underlying host specialization (in particular candidate effector genes) were
343 found to exist as gene families in the genome of a number of plant pathogens
344 (Plissonneau *et al.* 2017). Gene families are created by gene duplication events, which
345 can lead to paralog formation. The presence of such paralogs can interfere with the
346 assessment of polymorphism (and by extension estimates of *d*N/*d*S ratios) as
347 sequencing reads originating from one paralog might map to another paralog and
348 generate erroneous variant calls. Ortholog identification analysis was performed on
349 coding sequences (CDS) extracted from *E.t. poae* gene models and from consensus
350 sequences of *E.t. typhina* and *E.t. clarkii*. Encoded protein sequences were translated
351 using transeq (EMBOSS 6.6.0.0; Rice *et al.* 2000) and analysed for
352 orthology/paralogy using OMA (version 2.1.1; default settings except MinSeqLen =

353 30; Altenhoff *et al.* 2015). OMA results were examined for 1-to-1, 1-to-many, many-
354 to-1 and many-to-many ortholog relationships.
355
356 ***Detection of positive selection driven by host specialization***
357 To further identify signatures of positive selection in *E.t. typhina* and *E.t. clarkii* we
358 computed ratios of non-synonymous (*d*N) and synonymous (*d*S) substitutions
359 (Goldman & Yang 1994). The ratio of the two parameters is indicative of the type of
360 selection that has acted on a gene, where *d*N/*d*S < 1 indicates purifying selection,
361 *d*N/*d*S > 1 indicates positive selection and *d*N/*d*S = 1 indicates neutral evolution. We
362 computed 90% majority consensus sequences from the Pool-Seq data for all genes
363 from each subspecies with a custom script. Polymorphisms with alternate allele
364 frequencies above the 90% threshold were considered as fixed between subspecies,
365 and shared polymorphisms with allele frequencies below the threshold in either
366 subspecies were excluded from the analyses. We calculated *d*N/*d*S ratios for 8,211
367 genes with the BUSTED method (Murrell *et al.* 2015) implemented in the Hyphy
368 package (Kosakovsky Pond *et al.* 2005). BUSTED provides a likelihood ratio test for
369 positive selection and reports a gene-wise *d*N/*d*S value and the probability (i.e. *p*-
370 value) of a gene to have experienced positive selection in at least one site and on at
371 least one branch.
372 We performed a GO enrichment analysis using Blast2Go (Conesa & Götz
373 2007) for 58 genes with *d*N/*d*S values significantly > 1. Significance of each
374 individual GO category was computed using Fisher's exact test with a significance
375 threshold of 5%. Correction for multiple testing was performed using a false
376 discovery rate (FDR) of 0.05. We tested for an enrichment of *d*N/*d*S values
377 significantly > 1 in 24 genes predicted to encode secreted proteins, and compared the

378 *d*N/*d*S values of the 57 $F_{ST}$-$D_{XY}$ outliers to the distribution of the rest of the aligned
379 genes. We then focused our downstream analyses on five candidate genes encoding
380 secreted proteins within the $F_{ST}$-$D_{XY}$ outliers, and tested for signatures of positive
381 selection using the MK test on http://mkt.uab.es/mkt/ (Egea *et al.* 2008). The MK test
382 compares the within and between species proportions of non-synonymous and
383 synonymous variation, and can provide convincing evidence on the type of selection
384 acting on specific candidate genes. An excess of non-synonymous substitutions
385 relative to non-synonymous polymorphisms suggests that a gene has been under
386 positive selection during species divergence and has therefore fixed more non-
387 synonymous mutations (McDonald & Kreitman 1991). A minority consensus
388 sequence was generated by calling polymorphic sites within each population using the
389 same parameters as in the PoPoolation analyses.
390
391 ***Targeted Sanger sequencing of candidate genes in additional isolates***
392 To further assess within subspecies variation in the five candidate genes encoding
393 secreted proteins, we collected sequence data from a larger population sampling to
394 provide evidence that signature of positive selection is persistent across the subspecies
395 distribution. To this end, we designed new primer pairs for PCR amplification using
396 the consensus sequences of *E.t. typhina* and *E.t. clarkii* in Geneious 6.1.8 (Table S2,
397 Supporting information) (Drummond *et al.* 2013). Targeted Sanger sequences were
398 produced for the same 40 individuals from the Aubonne site included in the Pool-Seq
399 data set, and a sampling of 21 *E.t. typhina* individuals and seven *E.t. clarkii*
400 individuals from different locations across Europe (Table S3, Supporting information).
401 We used the Sanger sequences obtained for the individuals originating from
402 the Aubonne site to validate allele frequencies from the Pool-Seq data. The data set

403 including only the European sampling was used to compute a multi-locus MK test
404 that we could compare to the signals of selection recovered in the Pool-Seq data from
405 the Aubonne individuals. Because primers were designed on outer exons of a coding
406 sequence, the Sanger sequences of the European sampling may lack a number of
407 nucleotides at the beginning or the end of a gene compared to data from whole
408 genome sequencing. To ensure that excluded nucleotides did not bias the results of the
409 MK test performed on the European sampling, we repeated the test on samples from
410 the Aubonne site using the individual Sanger sequences and compared the results to
411 those obtained from the Pool-Seq approach.

412 DNA of the European samples was extracted following the procedure reported
413 above. For all samples, PCRs were performed in 15 µl reaction volumes containing 1
414 µl DNA, 0.075 µl GoTaq Polymerase (Promega), 3 µl buffer, 10 mM of each primer,
415 2.5 mM dNTPs, and 25 mM $MgCl_2$. An initial polymerase activation step of 3 min at
416 94°C was followed by 35 cycles of 30 s at 94°C, 30 s at 55°C and 1 min at 72°C, and
417 a final step of 7 min at 72°C. PCR reactions were sequenced using BigDye
418 Terminator v. 3.1 on a 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA,
419 USA). Forward and reverse strands were assembled in Geneious v. 6.1.8 (Drummond
420 *et al.* 2013), and consensus sequences were aligned using the default plugin of the
421 software. The obtained nucleotide sequences have been submitted to GenBank under
422 accessions KU566808 - KU567105 (Table S4, Supporting information).

423

424 **Results**

425 ***Population genomics sequencing and read mapping***

426 Sequencing of pooled DNA resulted in 38.4 million paired-end reads for *E.t. typhina*
427 and 30.6 million paired-end reads for *E.t. clarkii* (Table S5, Supporting information),

428 corresponding to a total of 8.5 Gb of sequence data. These data sets include sequences
429 from both the endophytes and the hosts. After adapter and quality filtering, we
430 retained 37.4 million paired-end reads from the *E.t. typhina* dataset (~97.4%) and 30.1
431 million paired-end reads from the *E.t. clarkii* dataset (~98.4%) with a median Phred-
432 score of 38 (Table S5, Supporting information). Approximately 21.4% and 16.9% of
433 all quality filtered reads from *E.t. typhina* and *E.t. clarkii*, respectively, mapped to the
434 reference genome of *E.t. poae*. Overall, the sequencing data could be mapped to 1911
435 (*E.t. typhina*) and 1875 (*E.t. clarkii*) contigs of a total of 2027 contigs in the reference
436 genome. The *E.t. typhina* and *E.t. clarkii* mappings respectively covered 79% and
437 78% of the *E.t. poae* genome with at least one read. Median genome-wide coverage
438 was 39X and 22X for *E.t. typhina* and *E.t. clarkii,* respectively, and 42X and 24X for
439 genes annotated in the reference genome. We observed an increased proportion of
440 alignment coverage in coding regions of *E.t. poae* in comparison to intergenic regions,
441 as expected for more conserved regions. In coding regions, 97% and 96% of the *E.t.*
442 *poae* reference genome was covered by at least one read for *E.t. typhina* and *E.t.*
443 *clarkii*, respectively.

444 To investigate the source of the unmapped reads, we mapped these to the
445 genome of the model grass *Brachypodium distachyon* Bd21-1 a (downloaded from the
446 *Brachypodium* Genome Database; http://www. brachypodium.org). *Brachypodium*
447 *distachyon* belongs to the same subfamily (Pooideae) as the endophyte host grasses *D.*
448 *glomerata* and *H. lanatus*. In total, 35% of the reads could be aligned, confirming that
449 a large proportion of the unmapped reads represented plant genomic DNA. We
450 hypothesize that the remaining unmapped reads comprised reads from regions specific
451 to *E.t. typhina* and *E.t. clarkii*, repeats, and plant regions specific to *D. glomerata* and
452 *H. lanatus*.

453

454 ***Genomic divergence of E.t. typhina and E.t. clarkii***

455 In the two-subspecies dataset for the $F_{ST}$ analyses, we identified 177,421 variable sites
456 within 8,206 mapped genes that fulfilled the coverage thresholds. Overall, the mean
457 number of variable sites per gene was 21.9 (mean gene length 1,400 bp). We
458 computed $F_{ST}$ and $D_{XY}$ values for all genes, and found genome-wide mean values of
459 0.53 and 0.008, respectively. Genes with $F_{ST}$ above the 95% quantile of the
460 distribution, corresponding to a threshold of 0.843, were considered as strongly
461 differentiated outliers (Fig. 2). These included 410 genes and 3.0% of all variable sites
462 in the data set. $D_{XY}$ outlier genes above the 95% quantile of the distribution,
463 corresponding to a threshold of 0.019, included 396 genes and 9.6% of variable sites
464 (Fig. 2). The overlap of the $F_{ST}$ and $D_{XY}$ outliers ($F_{ST}$-$D_{XY}$; Fig. 2) included 57 genes,
465 which formed the primary set of candidate genes under divergent selection in *E.t.*
466 *typhina* and *E.t. clarkii* (Table S6, Supporting information).

467 To seek evidence of positive selection on nucleotide sequences within each of
468 the two subspecies, we computed the summary statistics π and Tajima's *D* of genetic
469 variation based on polymorphisms within *E.t. typhina* and *E.t. clarkii*, respectively.
470 Interestingly, we found a remarkable difference in the level of genetic variation in the
471 two *Epichloë* subspecies. In *E.t. typhina* we identified 124,896 SNPs and in *E.t.*
472 *clarkii* we identified 41,523 SNPs within 8,814 mapped genes (mean number of SNPs
473 per gene = 14.2 and 4.1, respectively). We used the SNP data to compute nucleotide
474 diversity (π) as 0.0039 in *E.t. typhina* and 0.0017 in *E.t. clarkii,* suggesting highly
475 different effective population sizes of the two endophyte subspecies (Table 1). When
476 comparing π between $F_{ST}$-$D_{XY}$ outliers and non-outliers, we found that π was
477 significantly reduced in $F_{ST}$-$D_{XY}$ outlier genes in *E.t. clarkii* compared to the rest of

478 the genome (two-tailed Wilcoxon rank-sum test; $p$ = 0.022). This finding supports a
479 scenario of positive selection acting on these loci during divergence of lineages, while
480 the reduced variation at the within-species-level reflects recent positive selection. In
481 *E.t. typhina,* this difference was not significant (two-tailed Wilcoxon rank-sum test; $p$
482 = 0.507; Table 1).

483 Another parameter that reflects the distribution of genetic variation within the
484 two subspecies is Tajima's *D*. Tajima's *D* was slightly negative in both *E.t. typhina*
485 and *E.t. clarkii,* with mean values of -0.514 and -0.308, respectively, indicating a
486 skew of the site frequency spectrum and an overall excess of low frequency alleles
487 compared to neutral expectations. In *E.t. typhina*, Tajima's *D* was significantly
488 reduced in genes within $F_{ST}$-$D_{XY}$ outliers compared to the rest of the genome (two-
489 tailed Wilcoxon rank-sum test; $p$ = 0.007; Table 1), indicating an excess of rare alleles
490 in this set of genes possibly reflecting past selective sweeps at the loci. In *E.t. clarkii*,
491 this difference was not significant (two-tailed Wilcoxon rank-sum test; $p$ = 0.358).
492

493 ***Prediction of gene functions***

494 To address the potential functional relevance of the $F_{ST}$-$D_{XY}$ outlier genes, we
495 conducted a functional annotation of the predicted gene sequences in the *Epichloë*
496 genomes. Gene ontology (GO) categories could be assigned to 5,669 (64%) of all
497 8,739 genes aligned between *E.t. typhina*, *E.t. clarkii* and the reference genome of *E.t.*
498 *poae*. Among the 57 $F_{ST}$-$D_{XY}$ outliers, GO categories could be assigned to 22 genes
499 (0.38% of genes with an assigned GO category). Among these, we found 19
500 categories to be significantly overrepresented (Fisher's exact test; $p < 0.01$; Table 2).
501 These categories could be assigned to three main biological processes: modification
502 of the cell wall (GO:0042545), secretion of proteins (GO:0009306) and catabolic

503 processes of xylan, a group of hemicelluloses found in plant cell walls (GO:0045493)
504 (Bastawde 1992). Possibly because of the low sample size of the $F_{ST}$-$D_{XY}$ outlier
505 genes, these GO categories were not significantly enriched after correction for a FDR
506 of 5% (Pawitan *et al.* 2005). This is a common result when the tested categories are
507 not independent; in this case one gene can have several GO categories (Clarke & Hall
508 2009).

509 Because we were mainly interested in genes involved in host specialization,
510 we focused further analyses on genes predicted to encode secreted proteins; among
511 the rest of the $F_{ST}$-$D_{XY}$ outliers, we did not predict functions that can be directly
512 associated with plant-fungus interactions. Overall, 624 of all 8206 genes that mapped
513 to the *E.t. poae* reference genome encoded signal peptides and were predicted to be
514 extra-cellularly secreted in *E.t. typhina and E.t. clarkii*. Among the 57 $F_{ST}$-$D_{XY}$
515 outliers, five genes were predicted to encode proteins secreted to the extracellular
516 space (Table 3). We considered these genes to be strong candidates involved in
517 divergent host specialization. Of these five candidate genes, two did not have any
518 similarity to proteins in databases (Table 4). The remaining three genes were similar
519 to genes essential for pathogenicity in fungal plant pathogens (Eaton *et al.* 2015; Rudd
520 *et al.* 2015). Two of these genes encoded enzymes that may be involved in the
521 degradation of cell walls (a carbohydrate esterase family 8 protein and an endo-1,4-
522 beta-xylanase; Eaton *et al.* 2015), and one encoded a chloroperoxidase with a putative
523 role in the suppression of host defence (Rudd *et al.* 2015).

524

525 ***Detection of orthologous and paralogous genes***

526 As genes underlying host specialization may belong to gene families subject to
527 duplication events (Plissonneau *et al.* 2017), we performed an orthology analysis

528 between the genes sequenced in *E.t. typhina and E.t. clarkii*. This analysis revealed 1-
529 to-1 relationships for 8650 genes, many-to-1 or 1-to-many relationships for 32 genes,
530 and many-to-many relationships for 38 genes. Eighty genes were shorter than 30
531 amino acids and could therefore not be analysed. All five candidate genes encoding
532 secreted proteins had 1-to-1 relationships suggesting that these genes have not
533 experienced duplication events. Fifty-five of the 57 $F_{ST}$-$D_{XY}$ outliers had a 1-to-1
534 relationship, while two genes had sequences shorter than 30 amino acids. We also
535 screened 25 genes with the highest *d*N/*d*S values (ranging from 4.435 to 9.749, see
536 below). Of these, 24 genes had a 1-to-1 relationship and one gene had a sequence
537 shorter than 30 amino acids. This indicates that our genes of interest that were of
538 sufficient length for the analyses were all orthologs in the *Epichloë* subspecies
539 investigated here.
540
541 ***Signatures of positive selection in E.t. typhina and E.t. clarkii***
542 We next addressed the signatures of positive selection in *E.t. typhina* and *E.t. clarkii*
543 by inferring and comparing the proportion of non-synonymous to synonymous
544 substitutions over all aligned genes. A list of the *d*N/*d*S ratios of all genes is available
545 as supplementary material (Table S7, Supporting information). Our automated
546 procedure for computing the 90% majority consensus sequences of the Pool-Seq data
547 identified 8,211 genes (93.2 % of all mapped genes of *E.t. typhina* and *E.t. clarkii*,
548 respectively) with a valid protein translation, and the dataset for the *d*N/*d*S analyses
549 included 174,916 fixed differences (i.e. substitutions) between *E.t. typhina* and *E.t.*
550 *clarkii* (mean number of substitutions per gene = 22.6). We tested the accuracy of the
551 automated alignment of the entire dataset by comparing *d*N/*d*S ratios to those
552 obtained from a subset of 15 manually aligned genes among $F_{ST}$ outliers encoding

553 secreted proteins. The correlation was highly significant (Spearman's rank-order
554 correlation; $\rho = 0.897$; $p < 0.001$; Fig. S1, Supporting information), confirming the
555 reliability of our alignment procedure.

556 The genome-wide mean *d*N/*d*S ratio of 8,211 genes between *E.t. typhina* and
557 *E.t. clarkii* was 0.424 (Fig. 3A). Among all genes, 58 genes with *d*N/*d*S ratios
558 significantly > 1 ($p < 0.05$; LRT) were identified, with a mean *d*N/*d*S ratio of 2.507,
559 and 32 genes encoded secreted proteins, with a mean *d*N/*d*S ratio of 2.453 (Table 3).
560 Among 410 genes within the 5% upper tail of the *d*N/*d*S distribution (*d*N/*d*S > 1.349),
561 58 genes encoded secreted proteins, with a mean *d*N/*d*S ratio of 2.328, and nine genes
562 encoded secreted proteins with *d*N/*d*S ratios significantly > 1 ($p > 0.05$; LRT), with a
563 mean *d*N/*d*S ratio of 3.340 (Table 3). Among all genes, we found the *d*N/*d*S ratios of
564 genes encoding secreted proteins significantly increased compared to genes encoding
565 non-secreted proteins (two-tailed Wilcoxon rank-sum test; $p < 0.001$; Fig. 3B). This is
566 in line with an enrichment of genes encoding secreted proteins with *d*N/*d*S ratios
567 significantly > 1 compared to genes encoding non-secreted proteins ($X^2$ test; $p <$
568 $0.001$; Table 5). The enrichment shows that secreted proteins-encoding genes are
569 evolving more rapidly than genes encoding non-secreted proteins. Among genes with
570 *d*N/*d*S ratios significantly > 1 ($p > 0.05$; LRT), we found 28 categories (including
571 0.49% of all genes with an assigned GO category) to be significantly overrepresented
572 (Fisher's exact test; $p < 0.05$; Table S8, Supporting information). These GO categories
573 were predicted to be involved in four main biological processes of lipoate biosynthetic
574 process (GO:0009107), urea catabolic process (GO:0043149), xylan catabolic process
575 (GO:0045493), and interaction with host via protein secretion (GO:0052051).
576 However, these were not significantly enriched when correcting for multiple testing
577 (FDR < 5%).

578 Within the 57 $F_{ST}$-$D_{XY}$ outliers, *d*N/*d*S ratios were significantly increased
579 compared to non-outlier genes, with a mean value of 0.95 (two-tailed Wilcoxon rank-
580 sum test; $p < 0.001$; Table 1). Three out of the five candidate genes (gene IDs:
581 477_41, 477_55, and 175_57) showed evidence of positive selection as indicated by
582 *d*N/*d*S ratios significantly > 1 and fell within the 5% upper tail of the *d*N/*d*S
583 distribution, whereas the other two genes (gene IDs: 572_15 and 1280_17) had *d*N/*d*S
584 ratios not significantly > 1 but were included in the upper range of the genome-wide
585 distribution (Fig. 3A; Table 4).

586 To further detect and infer the rate of positive selection within the two
587 subspecies on the molecular level, we conducted a MK test. While the *d*N/*d*S ratios
588 are based on fixed non-synonymous and synonymous substitutions of comparative
589 data between *E.t. typhina* and *E.t. clarkii,* the MK test contrasts polymorphisms
590 within one subspecies to substitutions between the two subspecies (Zhai *et al.* 2009).
591 The MK test revealed an excess of fixed non-synonymous substitutions compared to
592 non-synonymous polymorphisms for the five $F_{ST}$-$D_{XY}$ candidate genes within *E.t.*
593 *typhina*, indicating positive selection. However, we only found statistical significance
594 for one of these genes (i.e. 175_57; $X^2$ test; $p < 0.01$; Table S9, Supporting
595 information); the low within-species level of polymorphism for the other four genes
596 provided little power for this statistical test (Bierne & Eyre-Walker 2004; Fay 2011).
597 To overcome this limitation, we employed the multi-locus MK test on the five
598 candidate genes, which allows the comparison of independent genomic regions in a
599 single statistic test, and increases the power to detect a significant signal (Egea *et al.*
600 2008). We found a significant excess of fixed non-synonymous substitutions
601 compared to non-synonymous polymorphisms in the five candidate genes within *E.t.*
602 *typhina* ($X^2$ test; $p < 0.01$; Table 6). This evidence, complementary to the *d*N/*d*S

603 analyses, further supports a scenario of positive selection acting on our five candidate
604 genes within *E.t. typhina.* On the other hand, no significant signal of selection was
605 detected when we conducted the MK test using polymorphism data from *E.t. clarkii*
606 (data not shown). We suggest that the low genetic variation within this subspecies
607 reduced the ability to detect signatures of selection based on within-population
608 variation.
609
610 ***Signatures of positive selection in additional European populations***
611 The signatures of positive selection in the five candidate genes encoding secreted
612 proteins suggest that these loci have played a role in the host specialization of *E.t.*
613 *typhina* and *E.t. clarkii* at the Aubonne site. To investigate whether a similar signature
614 of selection could be identified in a broader sampling of the two subspecies we set out
615 to characterize genetic variation in the five genes in additional *E.t. typhina* and *E.t.*
616 *clarkii* isolates. To this end, we designed primers for PCR amplification of the five
617 candidate genes from representative samples of the European distribution of *E.t.*
618 *typhina* and *E.t. clarkii*. We obtained complete PCR products for four genes (Table S2,
619 Supporting information). For the fifth gene (gene ID: 572_15) we could only amplify
620 a fragment of the sequence, thus the locus was excluded from further analyses. The
621 four-gene data set included both isolates from the Aubonne site and from a larger
622 European sampling. The Aubonne data set included sequences for all 40 individuals
623 and comprised 149 variable sites. We calculated allele frequencies for both the *E.t.*
624 *typhina* and *E.t. clarkii* subspecies, and tested the correlation with allele frequencies
625 inferred from the Pool-Seq data. The correlation was positive and highly significant
626 (Spearman's rank-order correlation $\rho = 0.983$; $p < 0.001$; Fig. S2, Supporting
627 information), and neither approach showed a tendency to yield higher or lower allele

628 frequency estimates, thus confirming the reliability of estimates from the Pool-Seq
629 data (Rellstab *et al.* 2013; Fracassetti *et al.* 2015). For the European sampling, we
630 obtained a total 111 of 112 expected sequences, comprising 173 variable sites. The
631 multi-locus MK test revealed signatures of selection consistent with findings from the
632 Aubonne site: significant positive selection was recovered within *E.t. typhina* ($X^2$ test;
633 $p < 0.05$; Table 6), whereas no signal was detected in *E.t. clarkii*. The multi-locus MK
634 test performed on the individual Sanger sequences from the Aubonne individuals
635 confirmed a significant signal of positive selection on the four candidate genes ($X^2$
636 test; $p < 0.01$; Table 6), as obtained from the Pool-Seq data, proving a negligible
637 effect of the variation in sequence length between the Sanger and the Pool-Seq data
638 sets.
639

640 **Discussion**

641 ***Detection of genes involved in host specialization***

642 In this study, we set out to explore the genetic basis of host specialization in fungal
643 endophytes using population genomic data of the sympatric *Epichloë typhina*
644 subspecies *E.t. typhina* and *E.t. clarkii*. The genome-wide distribution of $F_{ST}$ between
645 *E.t. typhina* and *E.t. clarkii* revealed high values of relative differentiation, indicative
646 of advanced stages of divergence between the studied populations. Alternatively, such
647 inflation could result from demographic processes that are known to affect this
648 summary statistics, *e.g.* population size contractions causing genome-wide loss of
649 diversity in either or both subspecies. The inference of outliers from such distribution
650 may be problematic, as loci included in the tail are more likely to be false positives
651 resulting from stochastic processes. As a measure of absolute divergence, $D_{XY}$ is
652 independent of within-population variation, and requires a longer divergence time to

653 acquire a signal compared to $F_{ST}$ (Cruickshank & Hahn 2014). In fact, between *E.t.*
654 *typhina* and *E.t. clarkii* $D_{XY}$ has a distribution centred at very low values, and genes in
655 the tail of the distribution constitute likely candidates experiencing reduced levels of
656 gene flow. Genes showing high levels of relative and absolute differentiation in
657 comparison to the rest of the genome, as determined by the overlap of outliers of the
658 $F_{ST}$ and $D_{XY}$ summary statistics, constituted our main candidates for host
659 specialization between *E.t. typhina* and *E.t. clarkii*. We identified 57 outlier genes,
660 and substantiated our findings with additional evidence of positive selection
661 underlying the process of divergence at these loci.

662 The distribution of the two studied subspecies includes populations of *E.t.*
663 *clarkii* often in proximity to the more widespread *E.t. typhina* (pers. observation).
664 Consequently, the $F_{ST}$-$D_{XY}$ outlier genes may represent loci that have differentiated in
665 sympatry because of divergent selection imposed by the two hosts and heterogeneous
666 levels of gene flow along the endophyte genomes. However, an alternative hypothesis
667 is that divergence and specialization to different hosts occurred after geographic
668 isolation, and present-day sympatric populations represent a case of secondary contact
669 between the subspecies. Dobzhansky-Muller (DM) incompatibilities are predicted to
670 develop during geographic isolation (Orr 1995), and these may be coupled with loci
671 under selection for host specialization upon secondary contact (Bierne *et al.* 2011). In
672 this scenario, loci involved in DM incompatibilities, though not under divergent
673 selection, may show high differentiation between the diverged subspecies, and appear
674 as outliers in genome scans. Moreover, outlier genes of both $F_{ST}$ and $D_{XY}$ statistics
675 may arise from allelic classes under balancing selection in the ancestral population,
676 and become sorted by random processes in the descendant lineages, without
677 necessarily being involved in the adaptation process (Guerrero & Hahn 2017). Indeed,

genomic signatures arising from these processes are difficult to disentangle from those linked to adaptation in the extant populations. Although we cannot rule out the possibility that these processes played a role, complementary analyses based on methods to detect positive selection and functional predictions support our hypothesis that our candidate genes have indeed experienced divergent selection due to their likely role in host-fungus interactions. In particular, π and Tajima's *D* were significantly reduced in the $F_{ST}$-$D_{XY}$ outlier genes in *E.t. clarkii* and *E.t. typhina*, respectively, while *d*N/*d*S ratios were significantly higher in $F_{ST}$-$D_{XY}$ outlier genes compared to the rest of the genes, supporting the role of positive selection in the evolution of coding sequences. Consistent strong signatures of positive selection recovered for five candidate genes encoding secreted proteins at the Aubonne site and in the European populations further suggest that these specific genes evolved under divergent selection on a broader evolutionary scale. Finally, comparisons between orthologous proteins suggested that three of our candidate genes may be involved in responses to plant defence mechanisms and in the degradation of cell walls (Eaton *et al.* 2015; Rudd *et al.* 2015), thus supporting the notion that they play a role in host interactions.

Our study is limited in the proportion of the genome that has been surveyed. The regions of *E.t. typhina* and *E.t. clarkii* genomes that could not be mapped to the *E.t. poae* reference genome were not analysed here. Furthermore, for 13% of the coding regions in our data set, we could not assign a functional prediction. It is also possible that we have left out important candidate genes by focusing solely on genes encoding secreted proteins, e.g. genes encoding secondary metabolites or small RNAs, or proteins secreted via non-conventional pathways (Weiberg *et al.* 2013). Further analyses using a specific reference genome with improved annotation will

703 likely reveal further genomic elements that are important for host specialization in
704 *Epichloë* endophytes.
705
706 ***Secreted proteins as determinants of host specialization***
707 The GO enrichment analyses of the 57 $F_{ST}$-$D_{XY}$ outlier genes indicated three main
708 biological processes putatively involved in the selective process for host
709 specialization, including cell wall modification, xylan catabolic processes, and
710 secreted proteins. Genes encoding secreted proteins are known to physically interact
711 with host molecules (Presti *et al.* 2015), thus they constitute primary candidates
712 involved in host specialization. Though GO categories were not significant after
713 correcting for multiple testing, enrichment of genes encoding secreted proteins with
714 high *d*N/*d*S ratios supports a role for these genes in host specialization. These results
715 are further supported by the involvement of genes with high *d*N/*d*S ratios in, inter alia,
716 xylan catabolic processes, and the interaction with host via protein secretion as
717 indicated by the GO enrichment analysis. Moreover, we found strong evidence for
718 positive selection acting on the coding sequences of five genes encoding secreted
719 proteins within the $F_{ST}$-$D_{XY}$ outliers, as the *d*N/*d*S ratios fell in the upper range of the
720 genome-wide distribution, and for three genes even above the 95% quantile with
721 *d*N/*d*S ratios significantly > 1. Additional evidence of positive selection was found in
722 *E.t. typhina* using the MK test in both the Aubonne individuals and a representative
723 European sampling. The lack of a significant signal in *E.t. clarkii* likely resulted from
724 the low levels of polymorphisms in this subspecies (Terauchi & Yoshida 2010).
725 Overall, this evidence points to a central role of secreted proteins in establishing host-
726 fungus interactions in an endophyte system that sterilizes its host for sexual
727 reproduction.

728 So far, our knowledge on the genetics underlying host specialization has
729 largely come from pathogenic and castrating systems with a sexual reproduction,
730 where secreted proteins are known to mediate interactions with the host, suppress host
731 defence responses, or manipulate host cell physiology (Rep 2005; Kamoun 2007;
732 Presti *et al.* 2015). Underlying genes are thus predicted to be primary targets of
733 selection imposed by the host in a co-evolutionary arms race between the two
734 interacting systems (Dawkins & Krebs 1979; Giraud *et al.* 2008; Stukenbrock 2013).
735 Selection on genes encoding secreted proteins has been studied in *Z. tritici*
736 (Stukenbrock *et al.* 2011) and *M. lychnidis-dioicae* (Badouin *et al.* 2017). In sexual
737 *Epichloë* species, the life cycle entails an intriguing sequence of interactions with the
738 host grasses. First, the interaction is symptomless during within-plant growth, and
739 then switches to antagonistic during sexual reproduction. Indeed, the sterilising effect
740 caused by the fungus on the plant inflorescence resembles that of *Microbotryum*,
741 which grows endophytically in its host until it reaches the bud meristems and anthers,
742 where pollen is replaced with fungal teliospores (Akhter & Antonovics 1999). Among
743 our five candidate genes, three presumably play a pathogenic role in response to plant
744 defense mechanisms or cell wall degradation (Eaton *et al.* 2015; Rudd *et al.* 2015).
745 Purely mutualistic systems such as mycorrhizal fungi possess a very restricted
746 repertoire of cell wall degrading enzymes (Martin *et al.* 2008; Tisserant *et al.* 2012),
747 and a comparative genomic study indicated few pathogenicity-related proteins in
748 endophytes compared to more aggressive pathogens (Gazis *et al.* 2016). However,
749 these may be crucial for highly specialized endophytes such as sexual *Epichloë*.
750 *Epichloë festucae* has been shown to possess genes encoding a variety of degrading
751 enzymes (Eaton *et al.* 2015) likely required to degrade cuticle and epidermal cell
752 walls of the host to facilitate an increased uptake of host derived nutrients when the

753 endophyte switches to proliferative external growth during stromata formation (Lam
754 *et al.* 1995; Eaton *et al.* 2011). Moreover, in this system, the upregulation of genes
755 encoding cell wall degrading enzymes was observed in antagonistic mutants disrupted
756 in key signalling genes for mutualism (Dupont *et al.* 2015). Degrading enzymes
757 therefore appear to be associated with the antagonistic relationship engaged by the
758 fungus during stroma formation. The derived position of highly specialized
759 mutualistic interactions within a larger clade of grass pathogens indicates that
760 mutualists have evolved from pathogenic ancestors (Clay & Schardl 2002), thus it can
761 be hypothesized that cell wall degrading enzymes of sexual *Epichloë* species were
762 retained from the ancestral repertoire to modify the formation of plant inflorescences
763 and enable stromata formation. The putative functions of three of our five candidate
764 genes suggests that these may be associated with the reproductive phase of the life
765 cycle in our system. Experimental evidence demonstrating at which stage the genes
766 are expressed and functional validation of their role is needed to test this hypothesis.
767

768 **Conclusions**

769 Host specialization is a fundamental process that underlies many associations between
770 microbes and their hosts. Understanding the co-evolutionary dynamics that shape the
771 complex interactions between hosts and microbes is of particular relevance to assess
772 the emergence of agriculturally important pathogens and for predicting movements of
773 disease-causing microbes to humans (Lips *et al.* 2006; Burokiene *et al.* 2015; Munck
774 *et al.* 2015). We leveraged on the life cycle of the fungus *Epichloë* to investigate the
775 genetic basis of host specialization between the sympatric populations of *E.t. typhina*
776 and *E.t. clarkii*. The reproductive model of these obligate sexual subspecies is
777 characterized by unconstrained dispersal of haploid gametes in a process similar to

778 pollination, followed by fertilization and the production of meiotic ascospores that are
779 transmitted to infect new hosts. The host imposes selection on the ability of hybrid
780 genotypes to infect and establish symbiosis, effectively enforcing an extrinsic
781 postzygotic barrier and thus conferring selection on loci underlying host
782 specialization. Lineage separation in such systems conform to a model of divergence
783 in the face of gene flow, and genome scans relying on measures of population
784 differentiation can be used to detect loci underlying host specialization. As the
785 reproductive model of sexual *Epichloë* is shared among many Ascomycetes, similar
786 analyses could be more broadly applied to identify the genetic determinants of co-
787 evolution between plants and fungi. However, similar approaches are unlikely to be
788 appropriate for obligate biothrophs where mating occurs within the host. In such
789 systems, host adaptation establishes a complete postzygotic barrier with subsequent
790 genome-wide divergence among host specialized lineages.

791 Consistent with the expectation that secreted proteins play a dominant role in
792 host specialization of plant pathogens, we found strong signatures of divergent
793 selection in five genes encoding secreted proteins. Strong positive selection in three of
794 these genes is indicative of crucial role in the specialization to distinct hosts. These
795 genes may play a role during the antagonistic phase of the infections, where stromata
796 are formed and the plant inflorescence is sterilized. The convergence on small
797 secreted proteins playing a major role in pathogenicity is a striking feature of all plant
798 pathogens across kingdoms (Kamoun 2007; Presti *et al.* 2015). Despite independent
799 evolutionary trajectories, nearly all plant pathogens have evolved a large complement
800 of small secreted proteins that interfere with the host immune system. Identifying the
801 molecular functions of pathogenicity-related proteins will provide a comprehensive
802 and mechanistic insight in the evolution of host specialization.

**Acknowledgements**

This study was funded by Swiss National Science Foundation grant 31003A_138479 (AL). Sequences and genomic data were generated at the Genetic Diversity Centre (ETH Zurich). C.L. Schardl provided the reference genome and access to the structural annotation. A. Widmer provided key support for the genomics work and scientific advice. We thank M.C. Fischer, M. Scharmann, N. Zemp, and J. Buckley for methodological advice and scientific discussions, C. Michel for laboratory assistance, and D. Barry for helpful comments on the manuscript. We acknowledge the Editor and four anonymous reviewers for helpful comments that greatly improved the manuscript.

**Data accessibility**

The datasets supporting the conclusions of this article are available in Sequence Read Archive SRA (SRR5571977; https://www.ncbi.nlm.nih.gov/sra), and in GenBank (KU566808; https://www.ncbi.nlm.nih.gov/genbank).

**Authors' contributions**

SF designed research, MKS preformed research, MKS and SZ analyzed data, MKS, SZ, DC, EHS, AL and SF interpreted results and wrote the manuscript.

825 **References**

826 Akhter S, Antonovics J (1999) Use of internal transcribed spacer primers and
827 fungicide treatments to study the anther-smut disease, *Microbotryum violaceum*
828 (= *Ustilago violacea*), of white campion *Silene alba* (= *Silene latifolia*).
829 *International Journal of Plant Sciences*, **160**, 1171–1176.
830 Altenhoff AM, Škunca N, Glover N *et al.* (2015) The OMA orthology database in
831 2015: function predictions, better plant support, synteny view and other
832 improvements. *Nucleic Acids Research*, **43**, D240–9.
833 Arnegard ME, McGee MD, Matthews B *et al.* (2014) Genetics of ecological
834 divergence during speciation. *Nature*, **511**, 307–311.
835 Badouin H, Gladieux P, Gouzy J *et al.* (2017) Widespread selective sweeps
836 throughout the genome of model plant pathogenic fungi and identification of
837 effector candidates. *Molecular Ecology*, **26**, 2041–2062.
838 Barrett LG, Thrall PH, Dodds PN *et al.* (2009) Diversity and evolution of effector loci
839 in natural populations of the plant pathogen *Melampsora lini*. *Molecular Biology*
840 *and Evolution*, **26**, 2499–2513.
841 Bastawde KB (1992) Xylan structure, microbial xylanases, and their mode of action.
842 *World Journal of Microbiology & Biotechnology*, **8**, 353–368.
843 Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid
844 substitution in *Drosophila*. *Molecular Biology and Evolution*, **21**, 1350–1360.
845 Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis:
846 why genome scans may fail to map local adaptation genes. *Molecular Ecology*,
847 **20**, 2044–2072.
848 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina
849 sequence data. *Bioinformatics*, **30**, 2114–2120.
850 Branco S, Gladieux P, Ellison CE *et al.* (2015) Genetic isolation between two recently
851 diverged populations of a symbiotic fungus. *Molecular Ecology*, **24**, 2747–2758.
852 Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive
853 radiation in African cichlid fish. *Nature*, **513**, 375–381.
854 Bronstein JL (2009) The evolution of facilitation and mutualism. *Journal of Ecology*,
855 **97**, 1160–1170.
856 Bultman TL, Leuchtmann A, Sullivan TJ, Dreyer AP (2011) Do *Botanophila* flies
857 provide reproductive isolation between two species of *Epichloë* fungi? A field
858 test. *New Phytologist*, **190**, 206–212.
859 Bultman TL, White JF Jr, Bowdish TI, Welch AM (1998) A new kind of mutualism
860 between fungi and insects. *Mycological Research*, **102**, 235–238.
861 Burokiene D, Prospero S, Jung E *et al.* (2015) Genetic population structure of the
862 invasive ash dieback pathogen *Hymenoscyphus fraxineus* in its expanding range.
863 *Biological Invasions*, **17**, 2743–2756.
864 Burri R (2017) Interpreting differentiation landscapes in the light of long-term linked
865 selection. *Evolution Letters*, **1**, 118–135.
866 Busby PE, Ridout M, Newcombe G (2016) Fungal endophytes: modifiers of plant
867 disease. *Plant Molecular Biology*, **90**, 645–655.
868 Clarke S, Hall P (2009) Robustness of multiple testing procedures against
869 dependence. *The Annals of Statistics*, **37**, 332–358.
870 Clay K, Schardl CL (2002) Evolutionary origins and ecological consequences of
871 endophyte symbiosis with grasses. *American Naturalist*, **160**, 99–127.
872 Conesa A, Götz S (2007) Blast2GO: A comprehensive suite for functional analysis in
873 plant genomics. *International Journal of Plant Genomics*, **2008**, 619832.
874 Conesa A, Götz S, García-Gómez JM *et al.* (2005) Blast2GO: a universal tool for

annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Craven KD, Hsiau P, Leuchtmann A, Hollin W, Schardl CL (2001) Multigene phylogeny of *Epichloë* species, fungal symbionts of grasses. *Annals of the Missouri Botanical Garden*, **88**, 14–34.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–274.

Dawkins R, Krebs JR (1979) Arms races between and within species. *Proceedings of the Royal Society B: Biological Sciences*, **205**, 489–511.

Drummond AJ, Ashton B, Buxton S *et al.* (2013) Geneious v6.1 created by Biomatters. *www.geneious.com*.

Dupont P-Y, Eaton CJ, Wargent JJ *et al.* (2015) Fungal endophyte infection of ryegrass reprograms host metabolism and alters development. *New Phytologist*, **208**, 1227–1240.

Eaton CJ, Cox MP, Scott B (2011) What triggers grass endophytes to switch from mutualism to pathogenism? *Plant Science*, **180**, 190–195.

Eaton CJ, Dupont P-Y, Solomon P *et al.* (2015) A core gene set describes the molecular basis of mutualism and antagonism in *Epichloë* spp. *Molecular Plant-Microbe Interactions*, **28**, 218–231.

Egea R, Casillas S, Barbadilla A (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, **36**, W157–62.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Ellison CE, Hall C, Kowbel D *et al.* (2011) Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 2831–2836.

Fay JC (2011) Weighing the evidence for adaptationat the molecular level. *Trends in Genetics*, **27**, 343–349.

Feder JL, Egan SP, Forbes AA (2012) Ecological adaptation and speciation: the evolutionary significance of habitat avoidance as a postzygotic reproductive barrier to gene flow. *International Journal of Ecology*, **2012**.

Fracassetti M, Griffin PC, Willi Y (2015) Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata*. *PloS One*, **10**, e0140462.

Ganley RJ, Brunsfeld SJ, Newcombe G (2004) A community of unknown, endophytic fungi in western white pine. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10107–10112.

Gazis R, Kuo A, Riley R *et al.* (2016) The genome of *Xylona heveae* provides a window into fungal endophytism. *Fungal Biology*, **120**, 26–42.

Giraud T (2006) Selection against migrant pathogens: the immigrant inviability barrier in pathogens. *Heredity*, **97**, 316–318.

Giraud T, Gladieux P, Gavrilets S (2010) Linking the emergence of fungal plant diseases with ecological speciation. *Trends in Ecology & Evolution*, **25**, 387–395.

Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME (2008) Speciation in fungi. *Fungal Genetics and Biology*, **45**, 791–802.

Giraud T, Villaréal LMMA, Austerlitz F, Le Gac M, Lavigne C (2006) Importance of the life cycle in sympatric host race formation and speciation of pathogens.

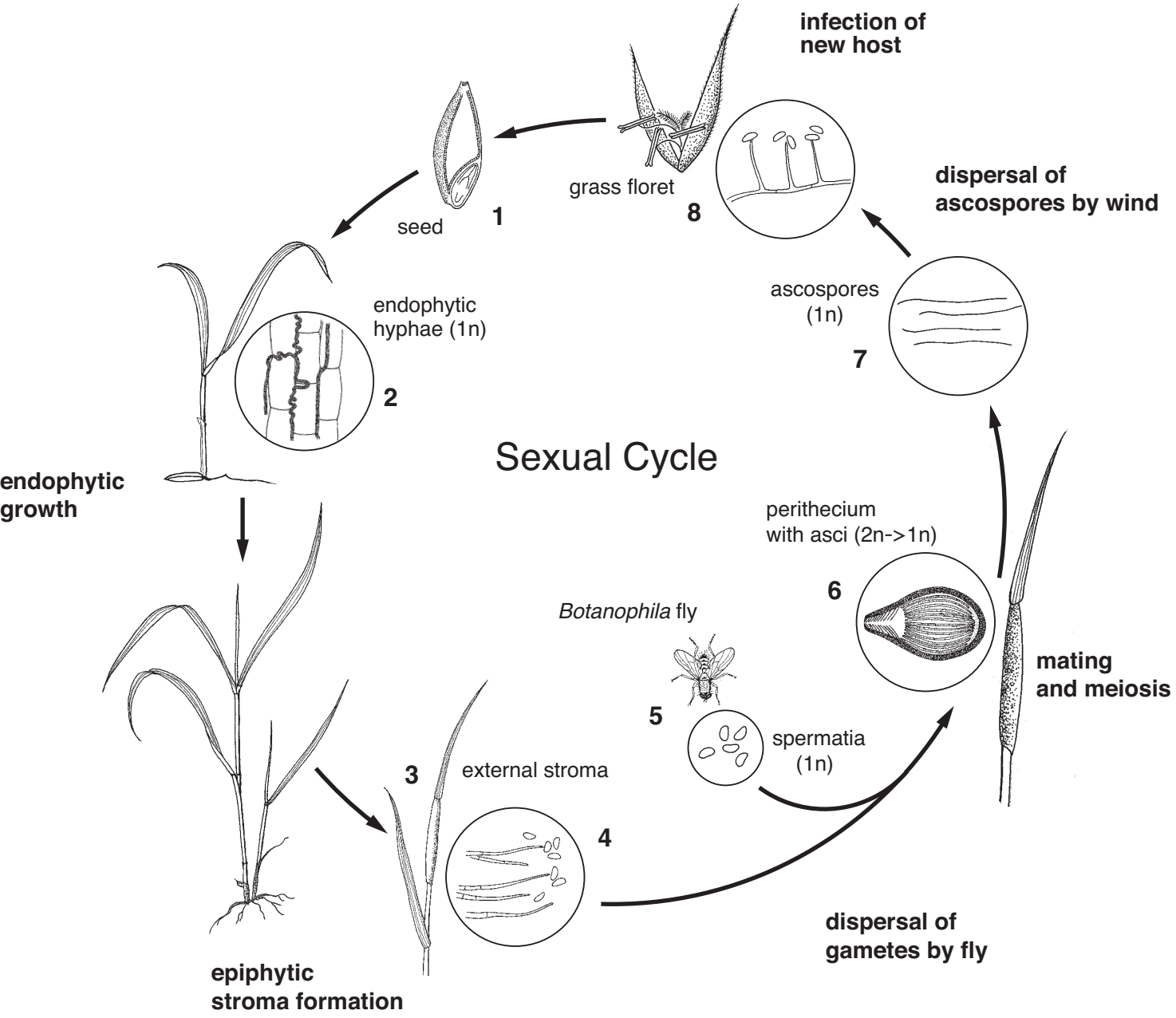
*Phytopathology*, **96**, 280–287.
Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725–736.
Guerrero RF, Hahn MW (2017) Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, **26**, 5362–5368.
Hartl DL, Clark AG (2007) *Principles of population genetics*. Sinauer Associates, Sunderland, MA, USA.
Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.
Jones P, Binns D, Chang H-Y *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
Kamoun S (2007) Groovy times: filamentous pathogen effectors revealed. *Current Opinion in Plant Biology*, **10**, 358–365.
Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, **338**, 1027–1036.
Kirby EJM (1961) Host-parasite relations in the choke disease of grasses. *Transactions of the British Mycological Society*, **44**, 493–503.
Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011a) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One*, **6**, e15925.
Kofler R, Pandey RV, Schlötterer C (2011b) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–3436.
Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.
Krogh A, Larsson B, Heijne von G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.
Lam CK, Belanger FC, White JF Jr, Daie J (1995) Invertase activity in *Epichloë*/*Acremonium* fungal endophytes and its possible role in choke disease. *Mycological Research*, **99**, 867–873.
Leuchtmann A, Schardl CL (1998) Mating compatibility and phylogenetic relationships among two new species of *Epichloë* and other congeneric European species. *Mycological Research*, **102**, 1169–1182.
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Lips KR, Brem F, Brenes R *et al.* (2006) Emerging infectious disease and the loss of biodiversity in a Neotropical amphibian community. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3165–3170.
Martin F, Aerts A, Ahrén D *et al.* (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature*, **452**, 88–92.
Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, **23**, 1817–1828.
McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.
Munck I, Livingston W, Lombard K *et al.* (2015) Extent and severity of caliciopsis canker in New England, USA: an emerging disease of eastern white pine (*Pinus*

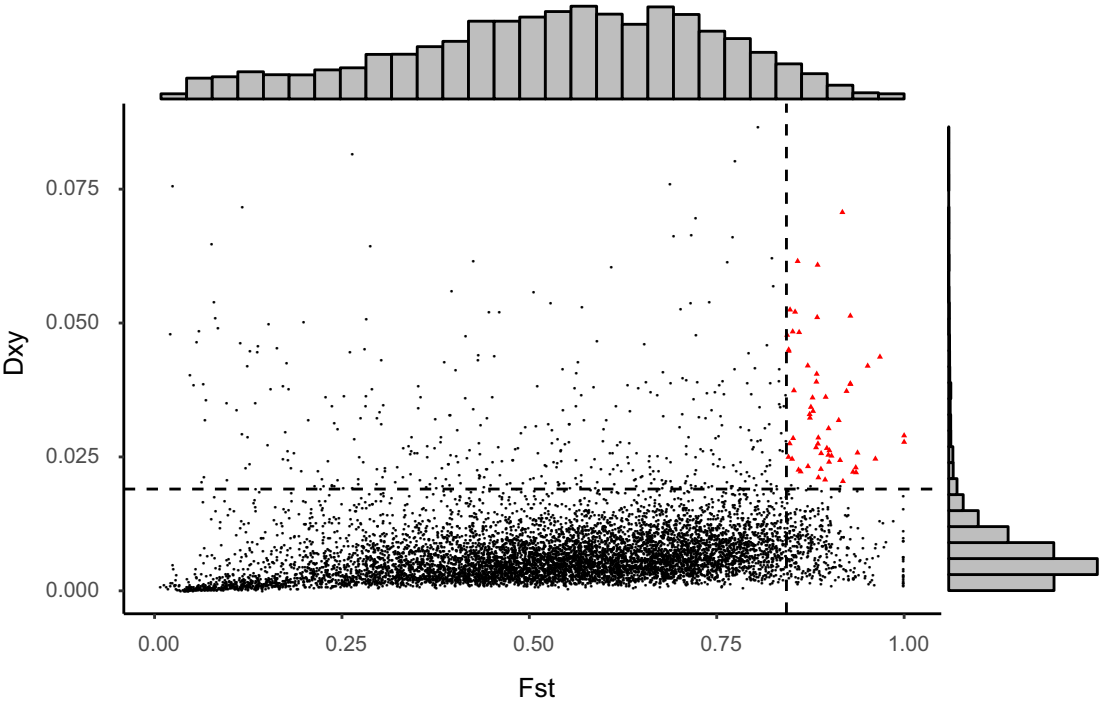
975 *strobus* L.). *Forests*, **6**, 4360–4373.
976 Murrell B, Weaver S, Smith MD *et al.* (2015) Gene-wide identification of episodic
977 selection. *Molecular Biology and Evolution*, **32**, 1365–1371.
978 Nachman MW, Payseur BA (2012) Recombination rate variation and speciation:
979 theoretical predictions and empirical results from rabbits and mice. *Philosophical*
980 *Transactions of the Royal Society B-Biological Sciences*, **367**, 409–421.
981 Nei M (1987) *Molecular evolutionary genetics*. Columbia University Press, New
982 York.
983 Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of
984 restriction endonucleases. *Proceedings of the National Academy of Sciences of*
985 *the United States of America*, **76**, 5269–5273.
986 Nosil P, Vines TH, Funk DJ (2005) Perspective: Reproductive isolation caused by
987 natural selection against immigrants from divergent habitats. *Evolution*, **59**, 705–
988 719.
989 Orr A (1995) The population genetics of speciation: the evolution of hybrid
990 incompatibilities. *Genetics*, **139**, 1805–1813.
991 Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A (2005) False discovery
992 rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–
993 3024.
994 Petersen TN, Brunak S, Heijne von G, Nielsen H (2011) SignalP 4.0: discriminating
995 signal peptides from transmembrane regions. *Nature Methods*, **8**, 785–786.
996 Plissonneau C, Benevenuto J, Mohd-Assaad N *et al.* (2017) Using population and
997 comparative genomics to understand the genetic basis of effector-driven fungal
998 pathogen evolution. *Frontiers in Plant Science*, **8**, 656.
999 Poelstra JW, Vijay N, Bossu CM *et al.* (2014) The genomic landscape underlying
1000 phenotypic integrity in the face of gene flow in crows. *Science*, **344**, 1410–1414.
1001 Poppe S, Dorsheimer L, Happel P, Stukenbrock EH (2015) Rapidly evolving genes
1002 are key players in host specialization and virulence of the fungal wheat pathogen
1003 *Zymoseptoria tritici* (*Mycosphaerella graminicola*). *PloS Pathogens*, **11**,
1004 e1005055.
1005 Presti Lo L, Lanver D, Schweizer G *et al.* (2015) Fungal effectors and plant
1006 susceptibility. *Annual Review of Plant Biology*, **66**, 513–545.
1007 Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial
1008 DNA: Contrasts among genes from *Drosophila*, mice, and humans. *Molecular*
1009 *Biology and Evolution*, **13**, 735–748.
1010 Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP
1011 allele frequencies determined by pooled next-generation sequencing in natural
1012 populations of a non-model plant species. *PloS One*, **8**, e80422.
1013 Rep M (2005) Small proteins of plant-pathogenic fungi secreted during host
1014 colonization. *FEMS Microbiology Letters*, **253**, 19–27.
1015 Restrepo S, Tabima JF, Mideros MF, Grünwald NJ, Matute DR (2014) Speciation in
1016 fungal and oomycete plant pathogens. *Annual Review of Phytopathology*, **52**,
1017 289–316.
1018 Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology
1019 open software suite. *Trends in Genetics*, **16**, 276–277.
1020 Rudd JJ, Kanyuka K, Hassani-Pak K *et al.* (2015) Transcriptome and metabolite
1021 profiling of the infection cycle of *Zymoseptoria tritici* on wheat reveals a biphasic
1022 interaction with plant immunity involving differential pathogen chromosomal
1023 contributions and a variation on the hemibiotrophic lifestyle definition. *Plant*
1024 *Physiology*, **167**, 1158–1185.

1025 Schardl CL, Leuchtmann A, Spiering MJ (2004) Symbioses of grasses with seedborne
1026 fungal endophytes. *Annual Review of Plant Biology*, **55**, 315–340.
1027 Schardl CL, Young CA, Hesse U *et al.* (2013) Plant-symbiotic fungi as chemical
1028 engineers: multi-genome analysis of the Clavicipitaceae reveals dynamics of
1029 alkaloid loci. *PLoS Genetics*, **9**, e1003323.
1030 Schirrmann MK, Leuchtmann A (2015) The role of host-specificity in the
1031 reproductive isolation of *Epichloë* endophytes revealed by reciprocal infections.
1032 *Fungal Ecology*, **15**, 29–38.
1033 Schirrmann MK, Zoller S, Fior S, Leuchtmann A (2015) Genetic evidence for
1034 reproductive isolation among sympatric *Epichloë* endophytes as inferred from
1035 newly developed microsatellite markers. *Microbial Ecology*, **70**, 51–60.
1036 Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU (2014) Combining
1037 experimental evolution with next-generation sequencing: a powerful tool to study
1038 adaptation from standing genetic variation. *Heredity*, **114**, 431–440.
1039 Smith J, Kronforst MR (2013) Do *Heliconius* butterfly species exchange mimicry
1040 alleles? *Biology Letters*, **9**, 20130503.
1041 Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick insect genomes reveal
1042 natural selection's role in parallel speciation. *Science*, **344**, 738–742.
1043 Stukenbrock EH (2013) Evolution, selection and isolation: a genomic view of
1044 speciation in fungal plant pathogens. *New Phytologist*, **199**, 895–907.
1045 Stukenbrock EH, Bataillon T, Dutheil JY *et al.* (2011) The making of a new
1046 pathogen: insights from comparative population genomics of the domesticated
1047 wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome*
1048 *Research*, **21**, 2157–2166.
1049 Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by
1050 DNA polymorphism. *Genetics*, **123**, 585–595.
1051 Terauchi R, Yoshida K (2010) Towards population genomics of effector-effector
1052 target interactions. *New Phytologist*, **187**, 929–939.
1053 Tisserant E, Kohler A, Dozolme-Seddas P *et al.* (2012) The transcriptome of the
1054 arbuscular mycorrhizal fungus *Glomus intraradices* (DAOM 197198) reveals
1055 functional tradeoffs in an obligate symbiont. *New Phytologist*, **193**, 755–769.
1056 Van Doren BM, Campagna L, Helm B *et al.* (2017) Correlated patterns of genetic
1057 diversity and differentiation across an avian family. *Molecular Ecology*, **26**,
1058 3982–3997.
1059 Vialle A, Feau N, Frey P, Bernier L, Hamelin RC (2013) Phylogenetic species
1060 recognition reveals host-specific lineages among poplar rust fungi. *Molecular*
1061 *Phylogenetics and Evolution*, **66**, 628–644.
1062 Vijay N, Bossu CM, Poelstra JW *et al.* (2016) Evolution of heterogeneous genome
1063 differentiation across multiple contact zones in a crow species complex. *Nature*
1064 *Communications*, **7**, 13195.
1065 Weiberg A, Wang M, Lin F-M *et al.* (2013) Fungal small RNAs suppress plant
1066 immunity by hijacking host RNA interference pathways. *Science*, **342**, 118–123.
1067 Western JH, Cavett JJ (1959) The choke disease of cocksfoot (*Dactylis glomerata*)
1068 caused by *Epichloë typhina* (Fr.) Tul. *Transactions of the British Mycological*
1069 *Society*, **42**, 298–307.
1070 White JF Jr, Bacon CW, Hinton DM (1997) Modifications of host cells and tissues by
1071 the biotrophic endophyte *Epichloë amarillans* (Clavicipitaceae; Ascomycotina).
1072 *Canadian Journal of Botany-Revue Canadienne De Botanique*, **75**, 1061–1069.
1073 Win J, Morgan W, Bos J *et al.* (2007) Adaptive evolution has targeted the C-terminal
1074 domain of the RXLR effectors of plant pathogenic oomycetes. *The Plant Cell*, **19**,

2349–2369.
Wolf JBW, Ellegren H (2016) Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, **18**, 87–100.
Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.
Zhai W, Nielsen R, Slatkin M (2009) An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution*, **26**, 273–283.

1084 **Figure legends**
1085 **Fig. 1:** Life cycle of *Epichloë* fungi. After systemic growth of haploid hyphae within
1086 seed (1) and vegetative plant tissues (2) sexual reproduction is initiated by forming an
1087 external fruiting body (stroma) around developing host inflorescences causing choke
1088 (3). On stroma surface, spermatia (male gametes) are produced (4) that are dispersed
1089 to stromata on other plants by *Botanophila* flies (5). Mating types prevent fertilization
1090 between spermatia and female structures on stromata from the same plant individual.
1091 Mating, karyogamy and meiosis take place on the fungal stroma (6). Ascospore
1092 progeny, which may be the result of mating within or between subspecies, are wind-
1093 dispersed (7) and mediate horizontal transmission to new hosts by infecting grass
1094 florets and then seeds (8). Figure modified from Leuchtmann & Schardl (1998).
1095
1096 **Fig. 2:** $F_{ST}$ values plotted against $D_{XY}$ values of all analysed genes in the genome. The
1097 horizontal line represents the threshold for the 5% quantile $F_{ST}$ outliers (> 0.843) and
1098 the vertical line the threshold for the 5% quantile $D_{XY}$ outliers (> 0.019). The overlap
1099 between $F_{ST}$ and $D_{XY}$ outliers is shown in the rectangle in the upper right. On the top
1100 of the x-axes is the frequency distribution of gene-wise $F_{ST}$ values and on the right of
1101 the y-axes the frequency distribution of genewise $D_{XY}$ values shown.
1102
1103 **Fig. 3: (A)** Frequency distribution of *d*N/*d*S ratios between *E.t. typhina* and *E.t.*
1104 *clarkii*. The 95% threshold for the positive *d*N/*d*S outliers (> 1.371) is shown with a
1105 solid vertical line. The *d*N/*d*S ratios of the five candidate genes are indicated by
1106 dashed lines. **(B)** Boxplots of *d*N/*d*S ratios of genes encoding non-secreted proteins
1107 and genes encoding secreted proteins. Asterisks indicate a significant difference
1108 between both categories (***$p < 0.001$).


infection of
new host
grass floret
8
seed
1
dispersal of
ascospores by wind
ascospores
(1n)
7
endophytic
hyphae (1n)
2
Sexual Cycle
endophytic
growth
perithecium
with asci (2n->1n)
6
Botanophila fly
5
spermatia
(1n)
mating
and meiosis
3
external stroma
4
dispersal of
gametes by fly
epiphytic
stroma formation


0.075
0.050
0.025
0.000
Dxy
0.00
0.25
0.50
0.75
1.00
Fst

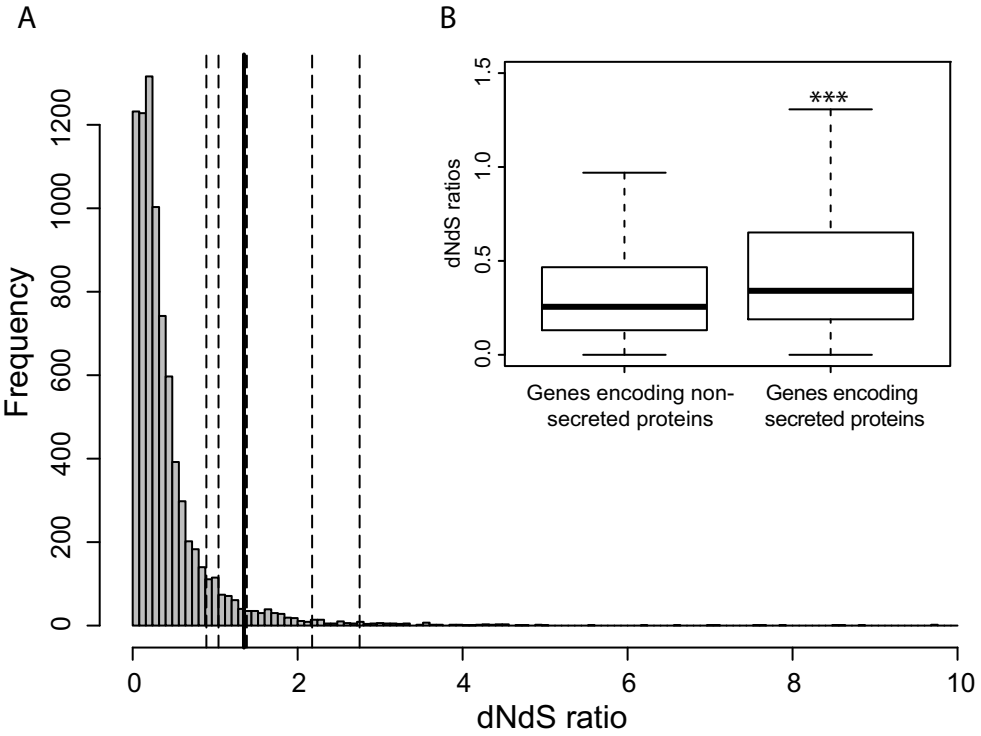

A
B
Frequency
0
200
400
600
800
1000
1200
0
2
4
6
8
10
dNdS ratio
dNdS ratios
0.0
0.5
1.0
1.5
***
Genes encoding non-secreted proteins
Genes encoding secreted proteins

**Table 1**: Population genetic summary statistics of $F_{ST}$-$D_{XY}$ outlier genes compared to non-outlier genes, including Tajima's *D* and $\pi$ within *E.t. typhina* and *E.t. clarkii,* and *d*N/*d*S ratios. For each statistic, mean values are shown. Asterisks indicate significant differences between $F_{ST}$-$D_{XY}$ outliers and non-outliers (***$p < 0.001$; **$p < 0.01$; *$p < 0.5$).

| Population | $F_{ST}$-$D_{XY}$ outliers | Non-outliers | *P*-value |
|---|---|---|---|
| $\pi$ | | | |
| *E.t. typhina* | 0.0035 | 0.0039 | 0.507 |
| *E.t. clarkii* | 0.0010 | 0.0017 | 0.022* |
| Tajima's *D* | | | |
| *E.t. typhina* | -0.865 | -0.514 | 0.007** |
| *E.t. clarkii* | -0.236 | -0.308 | 0.358 |
| *d*N/*d*S | | | |
| Inter-population | 0.948 | 0.405 | 1.538e-09*** |

**Table 2:** The 19 enriched GO categories significantly overrepresented before multiple testing (Fischer's exact test: $p < 0.01$) among $F_{ST}$-$D_{XY}$ outliers in *E.t. typhina* and *E.t. clarkii*.

| GO ID | GO category | *P*-value |
|---|---|---|
| GO:0071554 | cell wall organization or biogenesis | 1.80E-04 |
| GO:0044036 | cell wall macromolecule metabolic process | 8.92E-04 |
| GO:0005618 | cell wall | 1.82E-03 |
| GO:0030312 | external encapsulating structure | 2.04E-03 |
| GO:0004713 | protein tyrosine kinase activity | 2.78E-03 |
| GO:0031176 | endo-1,4-beta-xylanase activity | 3.81E-03 |
| GO:0019028 | viral capsid | 3.81E-03 |
| GO:0019013 | viral nucleocapsid | 3.81E-03 |
| GO:0030599 | pectinesterase activity | 3.81E-03 |
| GO:0042545 | cell wall modification | 3.81E-03 |
| GO:0006807 | nitrogen compound metabolic process | 3.89E-03 |
| GO:0005976 | polysaccharide metabolic process | 4.92E-03 |
| GO:1901e360 | organic cyclic compound metabolic process | 7.22E-03 |
| GO:0045493 | xylan catabolic process | 7.61E-03 |
| GO:0045491 | xylan metabolic process | 7.61E-03 |
| GO:0009306 | protein secretion | 7.61E-03 |
| GO:0004114 | 3',5'-cyclic-nucleotide phosphodiesterase activity | 7.61E-03 |
| GO:0004112 | cyclic-nucleotide phosphodiesterase activity | 7.61E-03 |
| GO:0010410 | hemicellulose metabolic process | 7.61E-03 |

**Table 3**: Number of all genes (# all) and of genes encoding for secreted proteins (# secreted) within the whole genome, $F_{ST}$-$D_{XY}$ outliers, genes with *d*N/*d*S ratios significantly > 1 (*d*N/*d*S significant), and genes within the 5% upper tail of the *d*N/*d*S distribution (*d*N/*d*S > 1.349).

| | # all | # secreted |
|---|---|---|
| Whole genome | 8206 | 624 |
| $F_{ST}$-$D_{XY}$ outliers | 57 | 5 |
| *d*N/*d*S significant | 58 | 32 |
| *d*N/*d*S > 1.349 | 410 | 58 |

**Table 4**: Population genetic summary statistics of $F_{ST}$-$D_{XY}$ outlier genes encoding for putative secreted proteins compared to non-outlier genes, including $F_{ST}$, $D_{XY}$, Tajima's $D$ and $\pi$, and $d$N/$d$S ratios. For candidate genes, the function is reported as well as $p$-values of likelihood ratio tests of the $d$N/$d$S ratios, the presence of a secretion signal, an extracellular domain, a transmembrane domain and a cytoplasmic domain.

| Gene ID | Gene name | Gene function | $F_{ST}$ | $D_{XY}$ | $D_{Ett}$ | $D_{Etc}$ | $\pi_{Ett}$ | $\pi_{Etc}$ | $d$N/$d$S | $P$-value | SS | EC | TD | CD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non-outliers | | 0.52 | 0.0074 | -0.514 | -0.308 | 0.0039 | 0.0017 | 0.394 | | | | | |
| | $F_{ST}$-$D_{XY}$ outliers | | 0.90 | 0.0252 | -0.865 | -0.236 | 0.0035 | 0.0010 | 0.920 | | | | | |
| 477_41 | maker-contig00477-fgenesh-gene-0.41 | Pectinesterase | 0.92 | 0.0373 | -1.757 | 0 | 0.0023 | 0 | 1.384 | 0.005** | + | + | + | – |
| 1280_17 | maker-contig01280-augustus-gene-0.17 | Peroxidase, family 2 (Chloroperoxidase) | 0.91 | 0.0318 | -0.852 | -1.133 | 0.0025 | 0.0003 | 0.894 | 1 | + | + | + | – |
| 175_57 | maker-contig00175-fgenesh-gene-0.57 | Glycosyl hydrolase family 10 (endo-1,4-beta-xylanase) | 0.88 | 0.0511 | -0.630 | 0 | 0.0085 | 0 | 2.751 | 0.001*** | + | + | + | – |
| 477_55 | maker-contig00477-augustus-gene-0.55 | NA | 0.86 | 0.0483 | 0.015 | 0.001 | 0.3080 | -0.4684 | 2.176 | 0.035* | + | + | –/+† | – |
| 572_15 | snap-masked-contig00572-processed-gene-0.15 | CVNH domain | 0.86 | 0.0615 | -1.293 | -0.808 | 0.0062 | 0.0032 | 1.040 | 0.235 | + | + | + | – |

$D_{Ett}$ – Tajima's $D$ *E.t. typhina*, $D_{Etc}$ – Tajima's $D$ *E.t. clarkii*, $\pi_{Ett}$ – $\pi$ *E.t. typhina*, $\pi_{Etc}$ – $\pi$ *E.t. clarkii*, CD – cytoplasmic domain; SS – secretion signal; EC – extracellular domain; TD – transmembrane domain; NA – not available; † transmembrane domain detected by TMHMM but not Phobius; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

**Table 5**: Number of genes encoding non-secreted and secreted proteins with *d*N/*d*S ratios significantly >1, 1 and <1 ($X^2$ = 41.653; $p$ = 9.021e-10***).

| | <1 | 1 | >1 |
|---|---|---|---|
| Non-secreted | 96 | 2 | 26 |
| Secreted | 4 | 0 | 32 |

**Table 6**: Results of the multi-locus McDonald-Kreitman test between *E.t. typhina* and *E.t. clarkii* for genes encoding for putative secreted proteins within $F_{ST}$-$D_{XY}$ outliers, and for individual sequences from Aubonne and Europe.

| Outlier | α | $\omega_{MH}$[a] | *P*-value |
|---|---|---|---|
| $F_{ST}$-$D_{XY}$ | 0.614 | 0.35 | 0.006** |
| Aubonne | 0.75 | 0.253 | 0.009** |
| Europe | 0.645 | 0.347 | 0.012* |

α – mean proportion of adaptive substitutions; $\omega_{MH}$ – Mantel-Haenszel estimator (equivalent to Neutrality Index; Rand & Kann 1996); **$p < 0.01$; *$p < 0.05$